西安交通大學

毕业设计（论文）

题　目　基于深度学习的工业图像异常检测算法设计与实现

电子与信息学部 学院 计算机科学与技术 专业 计算机少 71 班

学生姓名　刘逸伦

学　　号　2150506021

指导教师　赵季中教授

设计所在单位　西安交通大学软件所

2021 年 6 月

西 安 交 通 大 学

系（专业） 0537 计算机科学与技术（少年班）
系（专业）主任 唐亚哲
批 准 日 期 2020-11-24

毕业设计（论文）任务书

电子与信息学部 学院 0537 计算机科学与技术（少年班）专业 计算机少 71 班 学生 刘逸伦

毕业设计(论文)课题 基于深度学习的工业图像异常检测算法设计与实现

毕业设计(论文)工作自 2020 年 10 月 9 日起至 2021 年 6 月 14 日止

毕业设计(论文)进行地点：西安交通大学兴庆校区软件小楼 A304

课题的背景、意义及培养目标

异常检测一直是机器学习中一个非常重要的子分支，尤其是在工业检测领域。随着工业的发展，传统的异常检测速度已经不能满足目前的需求，更快地检测数据中的异常情况成为了当下非常重要的任务。工业检测缺陷检测算法利用图像处理和深度学习算法对某些部件图像进行分析，通过区分正常图像与缺陷图像之间的的特征，用来区分缺陷图像。由于缺陷图像数据与正常图像数据存在很大的不均衡现象，那么如何解决数据不均衡并准确的检测出缺陷图像，是目前存在的一个难题。

设计(论文)的原始数据与资料

1、图像处理相关知识；

2、深度学习相关知识；

3、图像分类及异常检测算法相关知识；

4、python 语言编程相关知识。

课题的主要任务

1、图像处理相关知识；

2、深度学习相关知识；

3、图像分类及异常检测算法相关知识；

4、python 语言编程相关知识。

课题的基本要求(工程设计类题应有技术经济分析要求)

1、能够实现对工业异常样本检测的基本功能；

2、对现有算法进行简单的改进，并提高检测的准确率；

3、实现在 windows 平台的 python 代码。

完成任务后提交的书面材料要求(图纸规格、数量，论文字数，外文翻译字数等)

1、15000 字左右的毕业设计论文一份；

2、英文翻译 3000 字左右；

3、掌握本课题国内外研究现状及发展趋势；

4、关键程序和算法。

主要参考文献

[1] Golan I, El-Yaniv R. Deep anomaly detection using geometric transformations[C] //Advances in Neural Information Processing Systems. 2018: 9758-9769.

[2] Perera P, Nallapati R, Xiang B. Ocgan: One-class novelty detection using gans with constrained latent representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2898-2906.

[3] Park H, Noh J, Ham B. Learning Memory-guided Normality for Anomaly Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14372-14381.

[4] Sabokrou M, Khalooei M, Adeli E. Self-supervised representation learning via neighborhood-relational encoding[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 8010-8019.

指导教师：赵季中

接受设计(论文)任务日期：2020-11-24

（注：由指导教师填写）　　学生签名：

西 安 交 通 大 学

毕业设计（论文）考核评议书

电子与信息学部 学院 0537 计算机科学与技术（少年班） 专业 计算机少 71 班

指导教师对学生 刘逸伦 所完成的课题为 基于深度学习的工业图像异常检测算法设计与实现

的毕业设计(论文)进行的情况，完成的质量及评分的意见： 论文针对缺陷具有类型多样、样本稀缺、尺度差异大等特点，研究基于小样本深度学习的工业异常检测算法。论文的研究方向关系国计民生，具有较好的发展前景。论文设计并实现了卷积神经网络和生成对抗网络融合的深度神经网络模型，可以通过生成器重建测试样本的正常图像模式、给出潜在异常区域的像素级掩膜，并基于此之上同时解决异常图像的分类、定位和分割问题。在 MNIST 手写体数字识别、CIFAR-10 自然图像以及 MVTecAD 工业异常检测数据集上对论文提出的算法进行的部署与测试。论文层结构合理，逻辑清晰，表达流畅，达到了本科毕业设计要求，同意答辩。

指导教师建议成绩： B-

指导教师 赵季中

2021 年 6 月 21 日

毕业设计（论文）评审意见书

评审意见： 论文针对图像表面缺陷检测，对基于生成对抗网络模型的异常图像分类、定位问题进行了一定研究，通过生成器重构测试样本的图像，获得异常区域的像素级掩膜并实现对缺陷类型的发现与定位，并达到较好的泛化能力和鲁棒性。论文第三章中 CNN 与 GAN 的基础知识部分应挪至第二章，论文实验部分应对损失函数及精度随迭代变化情况进行描述与分析。论文层次结构安排较为合理，表述较为清晰，文字流畅，基本达到了本科毕业设计（论文）要求。

评阅人建议成绩： B-

评阅人 赵鲲 职称 其他

2021 年 6 月 21 日

毕业设计（论文）答辩结果

电子与信息学部 学院

0537 计算机科学与技术（少年班） 专业

毕业设计(论文)答辩组对学生 刘逸伦 所完成的课题为 基于深度学习的工业图像异常检测算法设计与实现

的毕业设计(论文)经过答辩，其意见为 论文基于深度学习的工业异常检测，设计并实现了卷积神经网络和生成对抗网络综合构成的主体模型，大量实验结果表明该方法能够有效分辨工业异常图像。毕业设计工作量适中，有一定的难度，撰写的论文结构较合理，反映作者掌握了专业基础理论和专业知识，具有一定的独立工作能力。答辩过程中论述较为清晰，回答问题正确。答辩委员会根据学位申请人提交的材料、评阅人的意见和答辩情况，并经投票表决，一致同意授予刘逸伦同学工学学士学位，并确定成绩为 B+。

并确定成绩为 B+

毕业设计(论文)答辩组负责人 赵季中

答辩组成员 何晖 惠维

丁菡 王鸽

赵鲲

2021 年 6 月 21 日

摘 要

伴随着近半个世纪以来信息技术的发展与其在工业生产系统中的普及，在今天的工业异常检测工作流程中，图像样本采集、传输、响应处理与统计等阶段已基本实现自动化，但判定决策过程依然普遍依赖于人工辨别的传统程序，浪费了大量的人力物力，同时伴随着误差大、效率低、响应慢、耦合性差等缺点。建立适应新时代发展的工业故障检测技术是学界和业界共同的需求和期待。近年来，以机器视觉为基础的异常图像检测装备已在各领域逐渐大规模部署，不少基于神经网络的算法也被广泛应用在各种工业场景中。利用这样的方法进行工业异常检测，拥有传统方法无可比拟的多方面优势，应用前景十分广阔。

本文综合了深度学习有关算法与图像异常检测有关技术，设计并实现了由卷积神经网络（CNN）和生成对抗网络（GAN）综合构成的深度神经网络工业异常检测算法系统主体模型，可以通过生成器重建测试样本的正常图像模式、给出潜在异常区域的像素级掩膜，并基于此之上同步解决对于异常图像的分类、定位和分割问题。相较于单独将 CNN 或 GAN 模型引入并应用到基于机器视觉的工业异常检测场景中，该算法综合提取了两个习用模型各自的优点并加以结合，且相应地遵循了对于小样本问题和背景噪声等图像异常检测难点的解决途径，能够有效地达成高可解释性、非平衡样本适应性、高召回率和低假阳率等系统控制要素，在一定程度上解决了目前已有部分算法的不足之处。

在 MNIST 手写体数字识别、CIFAR-10 自然图像以及 MVTecAD 工业异常检测数据集上对本文算法进行的部署、测试和结果评估表明，其异常分类任务的 F1 指标（平均值）分别为 98.76%、89.16% 和 79.45%，且均能够生成较高质量的重建图像和异常区域掩膜，可以在异常检测任务中提供一致的有效结果。

关 键 词：异常检测；机器视觉；深度学习；卷积神经网络；生成对抗网络

ABSTRACT

With the development of information technology and its popularization in industrial production systems during the past half century, in today's industrial anomaly detection workflow, the stages of image sample collection, transmission, response processing and archiving have basically achieved automation; while the decision-making process still generally relies on the traditional procedures of manual identification, which wastes much manpower and material resources, accompanied by shortcomings such as large errors, low efficiency, slow response, and poor coupling. Establishing industrial anomaly detection technology that adapts to the development of the new era is the common demand and expectation of academia and industry. In recent years, Computer Vision-based image anomaly detection equipments have been gradually deployed on a large scale among various fields, and a number of Neural Network-based algorithms have also been widely used in various industrial scenarios. The use of such a method for industrial anomaly detection has many advantages over traditional methods and broad application prospects.

This paper integrates related Deep Learning algorithms and image anomaly detection technologies, and designs and implements the main model of the deep neural network industrial anomaly detection algorithm system composed of Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN). In the model of the algorithm, the generator can be used to reconstruct the normal image pattern of the test sample, which could be used to generate a pixel-level mask of the potential anomaly area, and based on which can the algorithm solve the problem of classification, location and segmentation of anomaly image simultageously. Compared with introducting and applicating CNN or GAN models based on computer vision individually into industrial anomaly detection scenes, this algorithm comprehensively extracts the respective advantages of the two conventional models and combines them, which accordingly follows the solution to the difficulties of image anomaly detection such as small sample problems and background noise, and can effectively achieve system control key elements such as high interpretability, unbalanced sample adaptability, high recall rate and low false positive rate, as to a certain extent would be able to solve the shortcomings in some of the existing algorithms.

The deployment, testing and result evaluation of the algorithm in this paper on the MNIST handwritten digit recognition dataset, CIFAR-10 natural image dataset and MVTecAD industrial anomaly detection dataset show that the F1 index (average value) of its anomaly classification task is 98.76%, 89.16%, and 79.45% respectively, and all of which can generate high-quality reconstructed images and masks of anomaly areas. The algorithm in this paper can provide consistent and effective performance in anomaly detection tasks.

KEY WORDS: Anomaly Detection; Computer Vision; Deep Learning; Convolutional Neural Network; Generative Adversarial Network

目 录

1 绪论

1.1 研究背景和意义

常言道:“眼睛是心灵的窗户。”作为人类观察和认识世界的重要手段，视觉信息对于人类的生产生活具有极其重要的意义。相关研究估计，人类的感觉、学习和认知活动中有 80% 至 85% 均来自于视觉系统[①]。随着世界范围内电子计算机、传感器、通信、信息存储和图像处理等有关技术的迅速发展和大规模应用，现代工业、医药、交通、国防等领域均愈发呈现出系统化、集成化、智能化的趋势，而其背后均离不开计算机视觉（Computer Vision）技术的支持。其中，异常检测（Anomaly Detection）作为涉及范围广阔的各种不同领域中的重要任务，被学界和业界普遍认为是计算机视觉所持续关注的一门颇具挑战性的研究方向。

近半个世纪以来，信息技术的发展与其在工业生产系统中的普及使得海量工业图像数据得以被广泛地采集和存储。在今天的规模化制造业部门所使用的异常检测工作流程中，图像样本采集、传输、响应处理与统计等阶段已基本实现自动化，但判定决策过程依然普遍依赖于终端设备展示和人工肉眼辨别的传统程序，造成了大量的设备与劳动力的浪费，同时存在着主观误差大、检测效率低等问题。

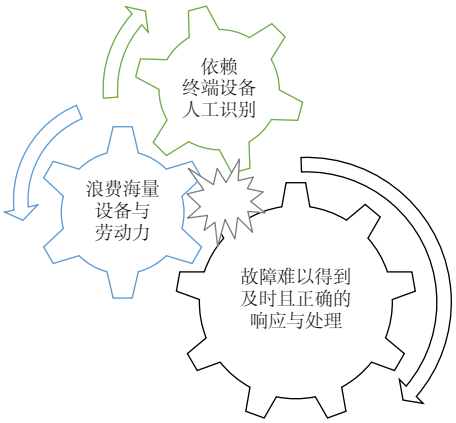


图 1-1　传统异常检测方法已不适应时代需要

此外，随着工业规模的扩大、复杂项目管理与控制部门的成熟以及设备精度要求的提升，各生产单元之间的耦合性升高，微小的故障在人工识别下常常难以得到及时且正确的响应与处理，容易在极短的时间内迅速逐级扩大问题规模，最终酿成严重的生产安全事故，产生巨大的经济损失。

① https://www.microsoft.com/en-us/research/project/projectflorence/

传统的人工工业异常检测方法已难以适应时代发展，建立更加适用于当今的工业故障检测技术是学界和业界共同的需求和期待。近年来，以机器视觉为基础的异常图像检测装备已在各领域逐渐大规模替代人工肉眼检测，不少基于神经网络的算法也被广泛应用在各种工业场景中。利用这样的方法来进行工业异常检测，拥有传统方法无可比拟的多方面优势：在提高响应速度和准确性的同时，降低了识别与处理成本、提升了产品质量、增加了产品的生产效益；同时，一体化的系统也提高了自动化系统的内聚性、降低了耦合，生产效率得到极大提升；此外，相关的算法模型不受具体生产部门的局限，可以很方便地推广并部署在相关行业乃至其他行业的各个生产环节中，应用前景十分广阔。

本文研究工作即基于上述有关背景展开，旨在将深度学习相关算法引入并应用到基于机器视觉的工业异常检测场景中，同时解决目前已有的部分算法的不足之处，使工业异常检测更加便捷、智能，提高相关部门的生产效益、效率和行业竞争力。

1.2 研究现状

1.2.1 广义异常检测研究现状

广义的异常检测任务是针对与正常样本相比被认为是异常的各种稀有样本、数据、物体或事件的识别。异常（anomaly）样本有时也称为离群样本：如果将每个样本视作样本空间中的一个点的话，异常样本即为其中距离大多数样本点都比较远的离群点，此时的异常检测任务即可视为在样本空间（例如向量空间、图像空间、特征空间等）中寻找正常样本与异常样本之间的界线的过程，同时尽可能将正常样本同异常样本更好地区分开来。

表 1-1 针对不同结构类型数据的异常检测应用领域和习用的机器学习模型[1]

结构类型	应用领域	习用模型
序列化数据	时间相干数据（语音、数值流）、自然语言文本等	RNN, LSTM 等
非序列化数据	空间相干数据（图像、传感器阵列数据等）	CNN, GAN 等
综合数据	时空相干数据（视频、传感器阵列流等）	CNN, RNN, LSTM 等

异常检测作为机器学习的重要任务，在众多相关领域中均有大量研究。表 1-1 给出了对于序列化和非序列化和时间相干、空间相干、时空相干等数据的不同的异常检测应用领域以及习用的机器学习模型，如循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）多用于时间相干数据，卷积神经网

络（Convolutional Neural Network, CNN）、生成对抗网络（Generative Adversarial Network, GAN）等则多用于空间相干数据。

基于部分关于异常检测方法的完整综述[1, 2]，对于各类异常检测任务而言，由于需要在工业环境下部署的特性，因此相较于其他机器学习任务，通常均要求模型应尽可能地具有更加高的召回率（Recall, true positive rate），通过避免遗漏真正的异常样本来确保系统能够正常运作；同时应当维持较低的假阳性率（Fall-out, false positive rate），不能将过高比例的正常样本误判定为异常，否则会耗费过多不必要的验证和处理成本。此外，当前学界与业界在异常检测中所普遍遇到的困难之一，是在真实环境（in the wild）下的小样本问题：相对于正常样本而言，异常样本往往呈现出数量少、难以获得的特点，甚至很多时候在前期模型训练过程中某些类型的异常样本没有采集，而在测试和部署时仍然要求模型能够正确检出这些缺失的异常类别。

对于简单的二分类任务，样本不均衡问题会影响模型输出。对于许多基于阈值的模型，其默认阈值为输出值的中位数，例如 Logistic 回归即以 0.5 作为反例与正例的分界阈值。此类模型的输出在数据出现不平衡情况时，往往会倾向于样本数据较多的类别，产生虚假的高准确率（Accuracy），导致分类失败。常见的例子是机场安全检查的恐怖分子判别任务：因为恐怖分子数量极少，因此即使模型将所有样本全部判为非恐怖分子也能拥有极高的准确率；然而该模型却完全没有能力真正识别出恐怖分子，无法在事实上达成目标任务。

传统的用来解决二分类任务下样本不均衡问题的方法可以归类于以下 3 个出发点：样本采样、模型改进和样本扩增。其中，样本采样分为过采样[4]（Oversampling）技术和欠采样[5]（Undersampling）技术两类，可以提升模型的泛化能力。过采样技术通过简单地产生大量重复样本的操作补齐少数类的样本，使得数据总量回到均衡状态，结果较为稳定；但并未向模型中引入更多实质性数据，同时这样的技术仅对单一类别的样本数据进行强调会将该类数据中的噪音的影响同步放大，进而导致分类器倾向于对正例的过拟合状态。欠采样技术则直接丢弃了部分大样本的反例数据，可能带来严重的信息丢失，从而导致模型学到的样本空间扭曲与不完整；不丢弃反例数据的欠采样技术则同过采样技术一样重复使用了正例样本，亦存在过拟合问题。两类模型采样的方法均可能改变原始数据空间的分布，从而引入系统性偏差。模型改进的方式从模型本身出发，常通过直接调整分类器的阈值与敏感度，或通过选择如 ROC（Receiver Operating Characteristic，受试者工作特征）曲线与 F1 指标等更合适的评估标准来替代准确度以提高诊断精度，往往提升程度有限，且通常较难获得最优的权重。样本扩增方法通过对负例样本进行分析并根据其分布合成虚拟的样本填充小样本空间，能有效避

免简单的采样方法存在的问题；然而，传统的样本扩增方法多运用诸如基于随机过采样算法的 SMOTE（Synthetic Minority Oversampling Technique，合成少数类过采样技术）及其改良模型[6]等，存在一定的局限性。

1.2.2 图像异常检测研究现状

在图像这一特定的异常检测方向上，对于正负样本的非平衡性问题，目前研究普遍采用的有 4 种不同的解决方式：

1、数据扩增与数据生成。传统的图像数据扩增方法通过在数据预处理阶段简单地在原始的异常样本之上使用镜像、旋转、平移、扭曲、滤波、对比度调整等多种类、多层次图像处理操作，来获取更多的样本补充数据集[8]；此外，也有部分研究将单独缺陷融合叠加到正常（无缺陷）样本上，以构成新的缺陷样本供后续模型进行学习[9]。这样的方法同前文所述的样本采样方法相类似，容易扩大小样本数据一侧的噪声，且可能导致过拟合现象的发生。

2、网络预训练与迁移学习。采用小样本来训练深度学习网络很容易导致过拟合，因此使用基于预训练网络或迁移学习的方法，可以通过引入另一相似且样本充裕的领域的数据先行训练，在一定程度上解决目标领域内样本的不平衡性问题[10]。相较于从零开始训练网络，此类方法的性能总是有所提高[11]；然而其局限性在于并非对于所有领域都存在其他相似且有大量样本可供使用的领域，且这样的迁移学习不能保证两个领域内数据空间分布的相似性，可能会引入额外的训练成本和系统性误差。

3、通过网络结构减少样本需求。可以引入专门的前置深度学习模型对小样本数据进行压缩和扩充以及提取特征[12]。例如，引入基于压缩采样定理的卷积神经网络（CNN）方法，可以在压缩采样后的数据特征空间进行后续的分类训练。相比于直接使用原始的图像空间作为输入，此类方法总是能够在少量样本上捕捉到更多可供模型进行学习的特征，在一定程度上降低对样本量的需求。然而，此类方法不能于实质上摆脱对于双方样本规模的依赖，若数据集规模过小则仍然容易产生过拟合情况。

4、采用半监督学习（semi-supervised）和无监督学习（unsupervised）的方法。在无监督模型中，只利用正常样本进行训练，在判定过程中发现未见过的异常特征时即判定为检测出异常，整个训练过程完全不需要异常样本的引入，直接摆脱了对于异常样本图像的需求；半监督方法则往往可以利用没有标注的样本来解决小样本情况下的网络训练难题[13]。

除了普遍存在的关于正负样本的非平衡性问题外，对于图像异常检测，通常还额外存在有图像的复杂背景噪声干扰，以及异常部位的尺寸、形态、位置不定所引起的类内差异大、类间差异小等技术难点。

1.3 本文研究内容

机器视觉任务中的“异常”这一语汇常常缺乏严格的数学定义，更多倾向于经验概念的范畴。不同的异常检测模型往往表现出对于异常的不同认知，例如，使用带类别标签、矩形定界框或像素级分辨率的掩膜等标签数据的异常图像的有监督方法通常关注和学习异常自身的特征，并以此作为异常判定标准；而只使用正常无缺陷样本的无监督异常检测方法则反其道而行之，标签在训练时是未知的，仅关注可供使用的正常样本特征，当判定过程中发现未见过的异常特征时即认为检测出异常。

与在机器视觉任务中有明确定义的分类（classification）、定位（localization）和分割（segmentation）任务相同，按模型功能、具体需求和目标应用场景可将异常检测任务简单划分为“异常是什么”（分类）、“异常在哪里”（定位）和“异常有多少”（分割）三个不同的层次，且三个层次间存在模式上的包含关系。现有的异常检测图像数据集所提供的标记模式不同，大部分均要求实现异常分类任务，而对于定位与分割仍有较大的研究空间。

本文研究工作即基于上述相关问题展开，专注于对工业领域基于无监督深度学习方法的、具有空间相干性的高维非序列数据（即图像）的异常检测进行研究，旨在将深度学习相关算法引入并应用到基于机器视觉的工业异常检测场景中，同时解决目前已有的部分算法的不足之处。本文首先以基于深度学习的工业异常检测算法为研究对象，开展了对异常检测相关算法和生成对抗网络模型的研究。根据相关技术背景和已有的机器学习与异常检测算法，本文设计并实现了基于卷积神经网络和生成对抗网络的算法模型，可以通过生成器重建测试样本的正常图像模式，给出潜在异常区域的像素级掩膜，并基于此同步解决对于异常图像的分类、定位和分割问题。本文在 MNIST 手写体数字识别、CIFAR-10 自然图像以及 MVTecAD 工业异常检测数据集上对算法进行了部署和测试，并评估了其综合性能。

1.4 本文组织结构

本文共分 5 个章节，按如下组织结构进行展开:

第 1 章，绪论。本章首先简要地阐述了基于深度学习的工业图像异常检测方法研究的背景及意义，对于广义异常检测和图像领域异常检测的难点、潜在解决方案和研究现状进行了简单的总结介绍。最后，阐明了本文的研究内容和结构安排。

第 2 章，相关研究概述。本章主要对本文相关的深度学习技术以及其在图像异常检测中的应用进行了概述，包括图像异常检测的研究框架和 GAN 的研究路径框架。

第 3 章，基于 CNN 和 GAN 的工业图像异常检测算法。本章首先简单介绍了基于

CNN、GAN 和工业异常检测算法的关键技术，之后分别对算法设计和系统结构进行了详细的阐述，包括算法的细节、目标函数的设计以及异常检测的完整过程。最后简要描述了算法实现。

第 4 章，实验与算法评估。本章首先介绍了使用的 MNIST 手写体数字识别、CIFAR-10 自然图像以及 MVTecAD 工业异常检测数据集以及针对本文环境的修改工作，随后对实验过程进行了简单介绍，并在数据集上围绕实验结果进行了详细的讨论与评估。

第 5 章，结论与展望。本章归纳总结了本文的主要研究工作，联系实际结果给出了基于深度学习的工业异常检测算法的应用价值和在实际中推广应用的可能性；在此基础上进一步围绕在本文研究中尚存在的问题和研究上的不足之处提出了见解与建议。

2 相关研究概述

2.1 图像异常检测研究

本节主要对深度学习的相关技术在图像异常检测中的应用进行了概述。图像异常检测的相关关键技术框架如下图 2-1 所示：

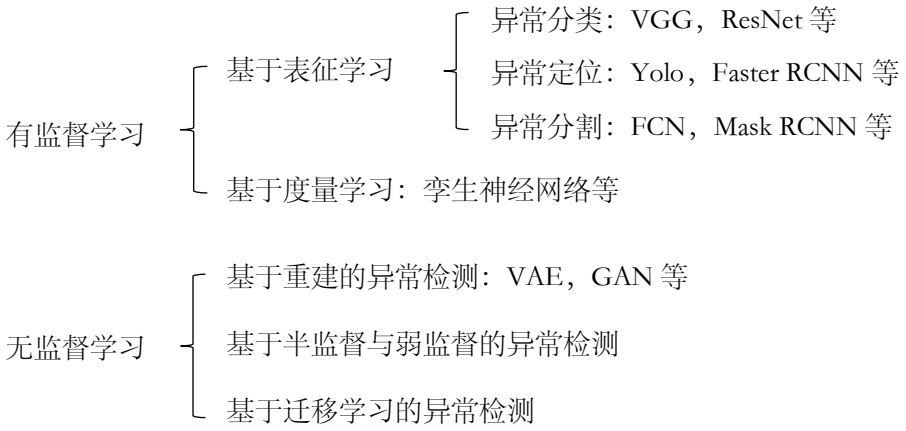


图 2-1 图像异常检测相关关键技术框架

本文专注于研究无监督深度学习中基于图像重建技术的工业图像异常检测。

2.1.1 有监督学习

1）基于表征学习的图像异常检测

表征学习的实质是将异常检测任务的分类、定位和分割问题全部视为计算机视觉中的分类任务，对应于图像标签分类、图像区域分类以及图像像素分类。现阶段已有大量基于有监督深度学习的表征学习的异常检测方法，大多均可以视为将相关的经典网络模型（多为 CNN）引入并应用至在工业图像异常检测领域。值得注意的是，过去对于使用 CNN 这一黑箱方法到异常检测中常常存在一些批评的声音，因为使用者对网络以简单的人类可理解的算法的形式所做的事情往往了解甚少；为此已有一定的 CNN 的可视化和可解释性工作于近年被提出并引起广泛讨论。

（1）异常分类

真实工业生产环境下的异常检测任务常囿于复杂环境中待检测样本的尺寸、形状、颜色、纹理质地、光照条件等因素的巨大差异而困难重重。基于 CNN（卷积神经网络）的网络模型结构，如 AlexNet[14]，VGG[15]，GoogLeNet[16]，ResNet[17]等，在这样的条件

下依然保持有强大的特征提取能力，是目前有监督的异常检测任务中最为通用的范式之一。基于此之上又有不同的分类子方法模型变种，包括直接分类、定位 ROI（Region of Interest，感兴趣区域）后分类、多类别分类，以及使用 CNN 做特征提取器、将特征传递给后续的其他机器学习分类器的网络。此外，值得注意的是，利用滑动窗口、热力图与多任务学习网络等技术，可以在分类的同时达成异常定位任务。

（2）异常定位

异常定位任务要求同时提供目标的类别信息和目标所在的精确位置（中心点、矩形定界框等）。从结构上，可以将目前已有的基于深度学习的异常定位网络大致分为以 YOLO（You Only Look Once）网络模型[18]为代表的，直接利用模型提取的特征预测异常的类别和位置的基于一阶段（one stage）的网络模型，和以 Faster R-CNN（Region-CNN）网络模型[19]为代表的，首先通过模型生成异常候选框（Proposal）、再进行目标定位的基于二阶段（two stage）的网络模型。符合直觉的是，基于一阶段的模型一般具有更快的检测速度，基于两阶段的模型则更多地被采用在需要强调检测精度的异常检测场景中。

（3）异常分割

异常分割任务可以被转化为正常与异常区域间在机器视觉上的语义分割与实例分割问题，要求对异常区域精细分割，同时包含异常的位置、类别以及相应的几何属性（包括长度、宽度、面积、轮廓、中心等）。可以按照分割功能的区别大致分为对应于语义分割的 FCN（Fully Convolutional Networks，全卷积神经网络）[20]方法和对应于实例分割的 Mask R-CNN 方法[21]。异常分割任务与异常定位类似，需要大量且逐像素的标注数据，往往耗费大量的标注精力和成本。

2）基于度量学习的图像异常检测

深度学习中的度量学习（Metric Learning）指的是一种空间映射的方法，学习到一种基于度量（距离函数）的特征（Embedding）空间，数据被转换成特征空间中的特征向量，且空间中相似样本与相互距离小的特征向量相对应。度量学习可以被近似地视为在特征空间对样本进行聚类：这与表征学习类似于在特征空间学习样本间分界的过程相不同。度量学习相关方法大多应用于异常分类任务中，在异常定位中应用较少。

异常分类任务的度量学习往往采用孪生网络[22]（Siamese network，又称为双生神经网络）进行度量学习，其通常以两个结构相同、权值共享的神经网络构成，使用两幅或多幅的成对图像作为输入，通过学习出输入样本的相似度，判断其属于同一类与否。孪生神经网络损失函数的核心思想即尽可能减小相似的样本间的距离、增大不同类别的

样本间的距离。在复杂的工业环境下，孪生神经网络要求输入图像统一内容形式，较为严格且不易达成。

2.1.2 无监督学习

异常检测任务中的无监督学习模型大多基于只使用正常无缺陷样本进行网络训练的单类别学习方法（One-class learning），对正常样本分布具有强大的学习、重建与判别能力。同有监督学习模型相比，只使用正常样本（在一些特定的弱监督工作[23]中，可用的正常样本集里可能有异常噪声样本掺杂）的无监督学习可以检测到偏离正常样本分布或从未在训练阶段出现过的异常模式：对于这样的异常输入样本，在重建和判别的过程中会产生同正常样本相异的结果。基于重建的无监督的异常检测的常见方法大致可分为基于图像空间和特征空间两种，多采用如 VAE[24]（Variational Autoencoder，变分自编码器）的 Autoencoder（自编码器）结构和 GAN（生成对抗网络，相关算法详细内容见 2.2 节生成对抗网络研究部分）结构作为网络模型。此外，也有部分研究基于半监督、弱监督以及基于迁移学习的算法和网络模型展开。

1）基于重建的图像异常检测

基于重建的图像异常检测方法不仅能实现图像标签级别的分类，同时提供了在图像中定位和分割异常的能力。这样的基于正常样本学习的重建方法可解释性好，且中间过程生成的重建图像在工业领域亦拥有后续的应用场景。目前，基于重建的方法多用于工业领域较为初级的结构、纹理等异常检测；在复杂的真实工业异常检测场景中，其检测效果与有监督学习的有关方法相比，较易受噪音影响和编码器和对抗网络模型的生成能力限制，其各方面能力有待后续研究进一步提升。

（1）基于图像空间重建的图像异常检测

类似于能够自动修复异常区域的去噪自编码器等模型，具有基于图像空间的样本重建与补全功能的网络模型算法[25]的生成器具有对于任意类型的样本图像输入，均输出其重建后得到的相应正常无缺陷样本图像的能力；而对于测试样本是否异常的判断指标则可以使用重建图像与输入图像进行相减、异或等运算得到的残差图像（重建误差），当其值大于某个阈值后即可将输入图像判定为异常样本，异常区域范围即为残差过大的区域；否则即判定为正常样本。同时，也有基于图像空间的模型直接使用生成对抗网络的判别器作为区分异常样本和正常样本的分类器，如 DCGAN[26]（Deep Convolutional GAN，深度卷积生成对抗网络）。有关生成对抗网络的相关算法详细内容见 2.2 节生成对抗网络研究部分。

（2）基于特征空间重建的图像异常检测

基于特征空间的重建方法目的是找到一个低维的特征空间并在该空间上重建正常样本；随后将输入图像样本投影至该特征空间内，通过特征空间距离算得异常分数（anomaly score）来衡量重建图像与测试图像特征分布之间的差异，进行异常检测：当异常分数高于阈值时即可判定为异常样本。对特征空间进行建模时，可以使用自动编码器[27]、变分自动编码器[28]或生成对抗网络[29]等网络模型。基于特征空间的方法在实现异常图像分类方面已有大量研究成果，而像素级别的异常精确分割则在近年间通过编码器和 GAN 模块进行了与图像空间检测方法类似的实现。

2）基于半监督与弱监督的图像异常检测

目前，关于半监督和弱监督学习的图像异常检测方法的研究和应用少于有监督学习和完全无监督学习。基于半监督的异常检测多使用少量有标记数据和大量未标记数据进行模型训练；基于弱监督的方法则围绕弱标签进行展开，如仅通过对图像类别标签进行学习来获取异常定位和异常分割结果[30]，或使用被异常污染的有噪声训练标签[23]。在真实的工业生产和异常检测环境中，异常对象往往相当罕见且难以标记，且获取无异常数据的过程需要将正常的数据标记出来，因此相比于完全有监督和无监督的方法，这样的深度学习模型往往更贴近实际场景。目前的方法多围绕异常类别进行展开，对于定位和分割仍有广阔的研究前景。

2.2 生成对抗网络研究

生成对抗网络（Generative Adversarial Network, GAN）最早由 Goodfellow I J 等人于 2014 年提出[31]，是一种基于博弈场景的半监督（或无监督）特征学习算法。随着对抗学习思想的不断完善，GAN 已经在图像生成、图像辨识和风格迁移等领域有了较多的应用，并且衍生出了实现不同功能的变体。生成对抗网络从样本生成的角度入手，训练生成器网络（Generator, $\mathcal{G}$）捕获真实数据的潜在分布、并且生成重建的数据样本，同时训练判别器网络（Discriminator, $\mathcal{D}$）鉴别输入数据为真实数据还是由生成器生成的样本，以此来指导生成器学习真实数据的分布。在训练过程中，GAN 并未通过计算公式概率求取数据真实分布，而是通过生成网络和鉴别网络的最大值-最小值（min-max）交替训练的博弈过程学习训练样本的数据分布，并且使用交叉熵（Jessen-Shannon divegence，JS 散度）计算两个分布的距离作为评估对真实数据建模的指标。根据学习到的真实分布，GAN 的生成网络能够输出以假乱真的重建图像样本，从而可以解决实际故障诊断中故障样本少于正常样本的数据不平衡问题。

2015 年提出的 DCGAN（Deep Convolutional GAN）网络[26]对 GAN 做出网络结构上的改进，取消了池化层和大量的全连接层，并参考卷积神经网络有关算法，引入分数步长（fractional-strided）卷积代替了上采样过程，在提取图像特征的同时增加了训练的稳定性。判别器和生成器几乎完全对称，使用步长卷积进行下采样，使得直接使用判别器即可作为区分异常样本和正常样本的分类器；不同之处在于，判别器使用 Leaky-ReLU 激活函数以防止梯度稀疏，而生成器中仍然采用 ReLU，且输出层采用 tanh 激活函数。DCGAN 为生成对抗网络提供了新的网络拓扑结构范式，同时最先表明了生成的特征具有向量的计算特性。目前，DCGAN 的网络结构仍在被广泛使用，极大的提升了 GAN 训练的稳定性以及生成结果质量。

WGAN（Wasserstein GAN）网络[32]与 DCGAN 不同，主要从损失函数的角度对 GAN 做了部分改进，易于实现且具有重大作用。WGAN 在理论方面给出了 GAN 训练不稳定的原因，即交叉熵等测距工具所具有的不连续性缺点，导致其不适合衡量生成数据分布和真实数据分布这两个具有不相交部分的分布之间的距离，使得鉴别器不能稳定地训练。为此，WGAN 提出使用 Wassertein 距离去衡量生成数据和真实数据分布之间的距离。使用 Wassertein 距离需要满足很强 Lipschitz 连续性条件，即要求判别器函数 $\mathcal{D}(x)$ 在样本空间中的梯度值不大于有限的常数 $\mathcal{K}$。WGAN 通过使用权重值限制的方式，强制保证了权重参数的有界性，间接限制了其梯度信息满足 Lipschitz 连续性。值得注意的是，WGAN 最先对 GAN 的训练提供了收敛程度的指标；此外，WGAN 还解决了模式崩溃（collapse mode）的问题，生成的结果多样性更加丰富。

WGAN-GP（improved WGAN，改进的 GAN）网络[33]改进了 WGAN 的连续性限制的条件，作者发现 WGAN 不能充分发挥深度神经网络的拟合能力，并且，也发现强制剪切权重容易导致梯度消失或者梯度爆炸。作者提出了使用梯度惩罚（gradient penalty）的方式，专门建立损失函数来满足 Lipschitz 限制的要求，从而解决了训练梯度消失梯度爆炸的问题。WGAN-GP 比标准 WGAN 拥有更快的收敛速度，并能生成更高质量的样本；此外还提供了稳定的 GAN 训练方式，成功训练多种针对图片生成和语言模型的 GAN 架构。

LSGAN（Least Squares GAN，最小二乘 GAN）网络[34]的基本原理是使用了最小二乘损失函数代替了 GAN 的损失函数，将图像的分布尽可能的接近决策边界，缓解了 GAN 训练不稳定和生成图像质量差多样性不足的问题。

由 Google 于 2017 年提出的 BEGAN（Boundary Equilibrium GAN）[35]是一种新的模型简单且功能强大的 GAN，判别器 $\mathcal{D}$ 使用了自动编码器（auto-encoder）结构。BEGAN 使用分布的误差来估计距离，代替了先前各种 GAN 设计损失函数来降低生成

数据分布与真实数据分布间的距离的方法。BEGAN 提供了又一标准的训练范式，能很快且稳定的收敛；且其使用不同的 γ 下的训练结果可以允许在图片的质量和生成多样性之间做选择。此外，基于 WGAN 的灵感，BEGAN 对于 GAN 中生成器和判别器的能力的平衡提出了估计收敛程度的又一衡量指标。

Google 于同年进行的一项研究[36]中比较了 GAN、WGAN、WGAN GP、LSGAN、BEGAN 以及 VAE 等模型在 FID（Fréchet distance，弗雷歇距离）、精度（precision）、召回率（recall）以及 F1 指标等方面的表现，结果表明，各个 GAN 在图像的生成质量方面表现相似。然而，其他各类的 GAN 无疑在训练数据量需求、训练速度、图像生成速度以及大尺寸图像生成等方面具有原始的 GAN 所不具备的优势。总体而言，WGAN-GP 在相关领域使用得更为广泛，而对于高清图像生成则更适合使用 BEGAN。

3 基于 CNN 和 GAN 的工业图像异常检测算法

综合考量系统设计难点及目前已有的相应解决方案，考虑到项目设计围绕深度学习方法展开,为了更好地解决工业图像异常检测问题,本文提出并实现了一种基于 CNN（卷积神经网络）和 GAN（生成对抗网络）的工业图像异常检测算法，同步解决对于异常图像的分类、定位和分割问题。

3.1 算法设计

传统的图像异常检测方法根据训练时所使用的样本数据标签类型，多将有监督学习范式下的 CNN 等技术同 GAN 等无监督学习模型割裂开来; 然而, 若能转变 CNN 受限于正负双方样本数量的需求的思路，将 CNN 强大的样本特征学习能力转而引入到 GAN 的重建图像生成机制中，则可以使得只使用正常样本进行训练的无监督学习模型捕捉到更加精确的正常样本模式，从而给出更加精确、可解释性更好的工业图像异常检测判别结果。

本文方法即结合卷积神经网络与生成对抗网络各自的优点，选取了针对特定难点的习用解决途径，提高了生成对抗网络对原始数据特征的挖掘能力，可以在异常检测任务中提供一致的有效结果。

参考 DCGAN 与 WGAN-GP 的基础架构，在由卷积神经网络和生成对抗网络结合构成的系统主体模型中，本文专门保留了 GAN 的生成器-判别器网络模型作为整体框架，同时引入了 CNN 的特征学习和表征机制:

1、在生成器中, 本文选择嵌入分数步长 (fractional-strided) 卷积层, 代替传统 CNN 中的池化层进行上采样；同时保留激活层，对图像数据进行分辨率扩增;

2、在判别器中，本文对图像使用步长卷积层和激活层进行有效的下采样，并最后传递给全连接层进行分类。

这样的模型架构可以在测试阶段通过生成器重建测试样本的正常图像模式、给出潜在异常区域的像素级掩膜，并基于此之上同步解决对于异常图像的分类、定位和分割问题。相较于单独使用 CNN 或 GAN 模型进行异常检测，该卷积神经网络（CNN）和生成对抗网络（GAN）综合构成的深度神经网络系统主体模型综合提取了两个习用模型各自的优点并加以结合, 能够有效地达成前述的高可解释性、非平衡样本适应性、高召回率和低假阳率等系统控制要素。详细的算法关键技术、系统架构和目标函数等将在接下来的章节内容中作进一步的描述。

3.2 算法关键技术

3.2.1 卷积神经网络

本文算法所使用的卷积层、反卷积层与激活层等相关技术，源自于卷积神经网络（Convolutional Neural Network, CNN）这一由卷积层、池化层和全连接层等堆叠形成的前馈多级神经网络。作为最具代表性的深度学习算法之一，自卷积神经网络受神经科学中的感受野（Receptive Field）机制启发而推出以来，在机器视觉的所有相关领域均已有大量的研究与应用。

传统的卷积神经网络由输入层、特征学习和表征层以及分类器三部分构成。

1、输入层将多维数据预处理成后续可用的结构；

2、特征学习和表征层由多个卷积层、激活层和池化层组成；

3、分类器通常由一个或几个全连接层组成，即多层感知器分类器，用于融合和分类提取特征。

其中，卷积层基于图像的局部关联性质（即像素作为高维非序列数据所具有的空间相干性）进行设计，是 CNN 的核心，也是本文算法所使用的卷积技术的核心部分。卷积原理如式 3-1 所示：

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{3-1}$$

式中：M_j——输入的神经元集合。卷积层包含多个卷积核，每个卷积核对应相应的权重矩阵和偏差矢量，类似于前馈神经网络的神经元，卷积核以预设的步长对给定的输入信号特征进行卷积操作，在感受野中获取被激活的局部特征进行学习，映射到 CNN 的隐空间（特征空间），将提取到的特征图输出，使得上一层的某个局部区域内的所有节点都被连接到下一层的一个节点上。

值得注意的是，在卷积层中，每个神经元仅对局部区域进行感知，而非以全连接的形式。这是因为全连接的形式拥有极多参数，需要大量数据和计算资源才能进行充分的前向和反向传播，在训练数据量不足时比较容易出现过拟合的情况。可以采用向卷积神经网络模型中加入 dropout 层的方法来防止过拟合情况的发生。此外，卷积层还具有权值共享特征：每个神经元对应的参数是相同的，同一个卷积核对不同数据进行处理时具有相同的权值和偏置值。这种局部连接与权值共享的卷积层网络特征，极大地降低了模型的参数量和运算量。

激活层一般使用 Sigmoid、Tanh（双曲正切）、ReLU（线性整流单元）、Leaky ReLU、ELU、Maxout 等常用的激活函数，对卷积层的结果进行非线性映射操作。池化层（Pooling）

对上层结果进行下采样，通过特征选择舍弃不必要的冗余信息、降低特征维度，同时保持大部分重要的信息，可以压缩数据量、减小过拟合，常用的有 Max Pooling（最大池化）和 Average Pooling（平均池化）两种操作。

网络尾部的全连接层对前一层的所有特征进行连接整合，所有神经元都有权重连接，来学习并完成最后的分类任务。

3.2.2 生成对抗网络

本文的系统主体结构围绕着生成对抗网络进行设计。生成对抗网络从样本生成的角度入手，训练生成器网络（Generator, $\mathcal{G}$）捕获真实数据的潜在分布、并且生成重建的数据样本，同时训练判别器网络（Discriminator, $\mathcal{D}$）鉴别输入数据为真实数据还是由生成器生成的样本，以此来指导生成器学习真实数据的分布。

典型的 GAN 优化目标函数式如下：

$$\min_{\theta_{\mathcal{G}}} \max_{\theta_{\mathcal{D}}} \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} q\big(\mathcal{D}(x)\big) + \mathop{\mathbb{E}}_{z\sim p_z} q\Big(1-\mathcal{D}\big(\mathcal{G}(z)\big)\Big) \tag{3-2}$$

式中：p_{data} ——真实数据样本的分布；p_z——生成器网络 $\mathcal{G}$ 所生成的样本分布。

生成器 $\mathcal{G}$ 以随机高斯噪声 $Z=(Z^1,Z^2,\dots,Z^m)$ 作为输入，通过反卷积的方式生成样本序列 $\mathcal{G}(z)=(\mathcal{G}(z)^1,\mathcal{G}(z)^2,\dots,\mathcal{G}(z)^{\mathrm{m}})$ 并输出给判别器进行判断，利用其判断结果反向训练生成器优化生成样本，优化目标是将生成的样本序列分布尽可能与真实样本分布相似，使得判别器网络 $\mathcal{D}$ 无法区分。单独训练生成器网络时，判别器网络模型参数被固定，优化函数为：

$$\min_{\theta_{\mathcal{G}}} \mathop{\mathbb{E}}_{z\sim p_z} q\Big(1-\mathcal{D}\big(\mathcal{G}(z)\big)\Big) \tag{3-3}$$

判别器网络 $\mathcal{D}$ 的目标是区分输入数据源自真实数据 X 还是生成器生成的样本序列 $\mathcal{G}(z)$。$\mathcal{D}$ 的输出对于 X 中的真实样本应尽可能接近 1，对于生成器网络 $\mathcal{G}$ 生成的样本则应尽可能接近 0。单独训练判别器网络时，生成器网络模型参数被固定。优化函数为：

$$\max_{\theta_{\mathcal{D}}} \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} q\big(\mathcal{D}(x)\big) + \mathop{\mathbb{E}}_{z\sim p_z} q\Big(1-\mathcal{D}\big(\mathcal{G}(z)\big)\Big) \tag{3-4}$$

无监督的生成对抗网络模型结构可以用来在训练时扩展潜在异常样本进行数据增强、重建正常模式以供判别差异、补足遮罩信息解决异常定位和分割问题，同时可以较为方便地迁移到不同的数据集上。

3.2.3 基于卷积神经网络与生成对抗网络的工业图像异常检测

对于工业异常检测的样本平衡性问题这一技术难点，采用卷积神经网络和生成对抗网络可以有效解决。

在数据扩增与数据生成方面，使用 GAN 可以用来在训练时扩展潜在异常样本进行数据增强，同时可以依靠生成器补足遮罩信息，以解决异常定位和分割问题。在网络预训练与迁移学习方面，CNN 的特征识别和 GAN 的生成器-判别器模型允许将预训练好的模型迁移至不同任务中。在通过网络结构减少样本需求方面，CNN 从网络结构入手，基于压缩采样的思想，将原始的图像输入转移到卷积操作后的数据特征空间来进行训练分类，可以使模型系统地学习到纹理、结构、空间等非像素类型的特征，能大大降低传统网络对海量样本的需求。在半监督学习与无监督学习方面，GAN 可以通过围绕正常样本进行模型训练，对测试图像尝试建立正常样本模式并判别差异来进行工业图像的异常检测任务。

图 3-1　基于卷积神经网络与生成对抗网络的工业异常检测相关关键技术

在本文所使用的网络模型中，围绕着卷积神经网络的样本特征感知特性和生成对抗网络的重建图像模式能力，重点从通过网络结构减少样本需求和无监督学习等方面入手，解决工业异常检测的样本不均衡问题等技术难点。

3.3 算法模型结构

本文在生成器和判别器的特征提取层，均使用了卷积神经网络（CNN）结构的特征学习和表征层，代替了原始 GAN 中的多层感知判别机制。为了在减小网络训练开支的同时保持整个网络结构可微，本文选择去除了传统 CNN 特征学习和表征层的卷积层、激活层和池化层三层结构中的池化层，只保留卷积层和激活层作为卷积和反卷积过程的核心。

3.3.1 算法细节

在以生成对抗网络为主要框架的整体架构中，生成器的输入是一个 128 维向量，随后首先使用了一个全连接层，将该向量同 4×4×4 倍生成器特征数量的节点相全连接，对应于第一层反卷积层 4 倍生成器特征数量的 4×4 像素分辨率图像的输入。而后使用三个反卷积层，分别使用大小为 5、5、8，插值为 1、1、2 的生成器特征数量个反卷积核，依次将图像分辨率由 4×4 逐级扩增至 8×8、12×12 和 28×28，对应的特征图像数量由 4 倍生成器特征数量逐级减半。最终，生成器输出一张 28×28 像素分辨率的图像作为生成器重建出的生成图像。

判别器的输入是一张 28×28 像素分辨率的图像，随后使用三个卷积层，分别使用大小均为 5，插值均为 2，填充均为 2 的判别器特征数量个卷积核，依次将图像分辨率由 28×28 逐级递减至 14×14、7×7 和 4×4，对应的图像数量由 1 张逐级翻倍至 1 倍、2 倍和 4 倍判别器特征数量。而后使用 1 个全连接层，将最后一个卷积层的 4×4×4 倍判别器特征数量的节点全链接至 1 个一维向量节点，作为判别器对输入图像是否属于生成器重建图像的最终判别结果输出。

本文所使用的生成器（上）和判别器（下）卷积网络模型结构详情如下图 3-2 所示，其中 ngf 为生成器特征数量、ndf 为判别器特征数量，默认值均为 64：

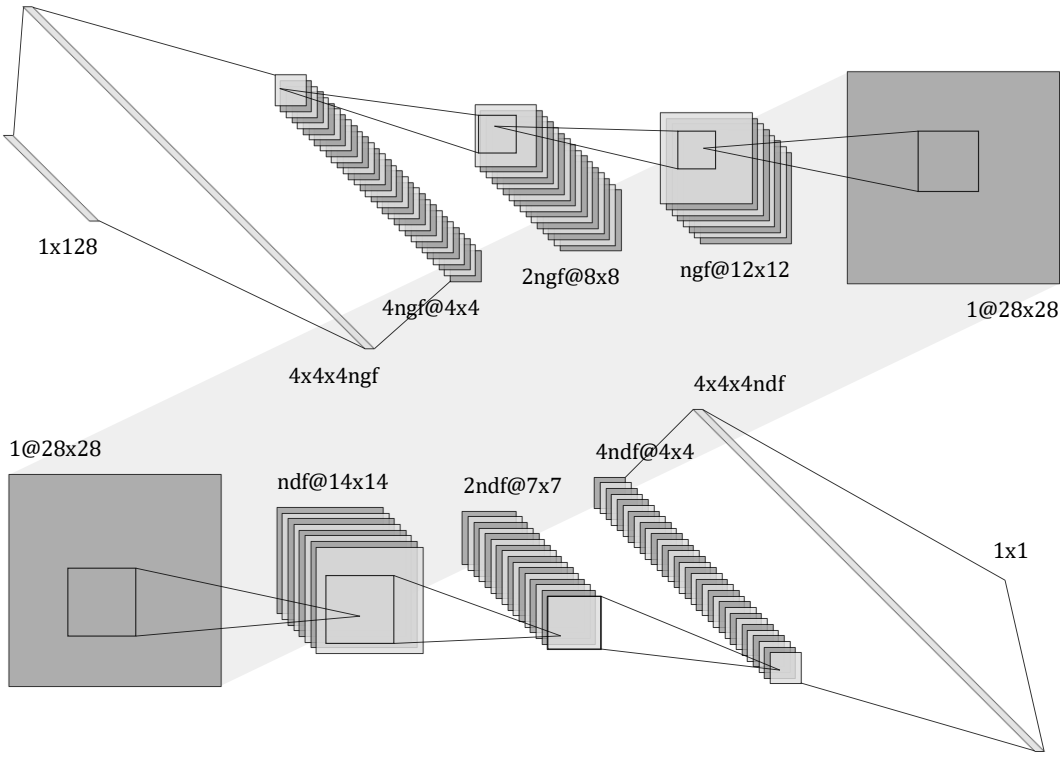


图 3-2　本文所使用的生成器（上）和判别器（下）卷积网络模型结构详图

在生成器和判别器的每一层卷积与反卷积过程后，均使用了 ReLU 线性整流作为激活层。其中，特别针对生成器的最后一层反卷积的输出结果使用了 Sigmoid 激活函数，以提高生成图像质量。

3.3.2 目标函数

本文算法将 $\mathcal{G}$ 与 $\mathcal{D}$ 联合训练。采用了 WGAN-GP 损失[33]，GAN 的目标函数为如下形式：

$$\mathcal{L} = \mathop{\mathbb{E}}_{z\sim p_z} q\left(\mathcal{D}\big(\mathcal{G}(z)\big)\right) - \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} q\big(\mathcal{D}(x)\big) + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim p_{\hat{x}}} q((\|\nabla_{\hat{x}}\mathcal{D}(\hat{x})\|_2 - 1)^2) \tag{3-5}$$

相较于原始的 WGAN 目标函数[32]：

$$\mathcal{L} = \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} q\big(\mathcal{D}(x)\big) - \mathop{\mathbb{E}}_{z\sim p_z} q\left(\mathcal{D}\big(\mathcal{G}(z)\big)\right) \tag{3-6}$$

式中第三项 $\lambda \mathop{\mathbb{E}}_{\hat{x}\sim p_{\hat{x}}} q((\|\nabla_{\hat{x}}\mathcal{D}(\hat{x})\|_2 - 1)^2)$ 是相较于 WGAN 所新提出的部分。由于使用 Wasserstein 损失[32]，故 $q(x) = x$；为满足 Lipschtiz 条件，使得可微函数梯度的范数处处至多为 1，WGAN-GP 使用了目标函数中的第三项约束 $\mathop{\mathbb{E}}_{\hat{x}\sim p_{\hat{x}}} q((\|\nabla_{\hat{x}}\mathcal{D}(\hat{x})\|_2 - 1)^2)$ 直接限制判别器输出结果对于输入值的梯度。为了便于处理，WGAN-GP 执行了更为宽泛的约束，直接对随机样本 $\hat{x}\sim p_{\hat{x}}$ 的梯度范数进行惩罚。

由 $\mathcal{G}$ 重建的待查询图像 $\mathcal{Q}$ 应当是在图像空间中距离待查询图像 $\mathcal{Q}$ 更近的匹配，而不是在隐空间中最接近的匹配；这是因为异常度指标 a（见下一节）正是部分基于图像空间中的距离度量所定义的。此外，图像空间损失以一种不同于隐空间损失的方式构造出了隐空间，从而将正常样本和异常样本分开。

3.3.3 异常检测

本文使用了由两个部分组成的异常度指标，即标准化残差（normalized residual）和原点距离损失（origin distance loss）。定义待查询图像 $\mathcal{Q} \in [0,1]^{W\times H\times D}$ 的残差损失 $\mathcal{L}_n$ 为 $\mathcal{Q}$ 与其最接近匹配 $\mathcal{G}(\hat{\mathbf{z}})$ 之间的 ℓ_2-范数：

$$\mathcal{L}_n\big(\mathcal{Q}, \mathcal{G}(\hat{z})\big) = \frac{1}{N_X}\left\|w(\mathcal{Q}) - w\big(\mathcal{G}(\hat{z})\big)\right\|_2 \tag{3-7}$$

式中：$\hat{z}$——重建图像 $\mathcal{Q}$ 所使用的向量。为了最小化图像对比度对残差损失的影响，本文采用图像的最小值－最大值标准化（minmax normalization）$w(x)$。定义标准化 $w(x): [\min(X), \max(X)]^{W\times H\times D} \mapsto [0,1]^{W\times H\times D}$ 为：

$$w(x) = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{3-8}$$

式中：$\min(X)$ 和 $\max(X)$——查找 X 中的最小值和最大值元素的过程；除法运算是在元素层面上（element-wise）进行的，且 $N_X = W \cdot H \cdot D$。若不进行最小值－最大值标准化，较低对比度的图像样本将具有较小的残差损失，反之亦然。

基于本文生成器和判别器交替的训练过程，本文将原点距离损失 $\mathcal{L}_{\mathrm{o}}$ 定义为向量空间中从 $\hat{\mathbf{z}}$ 到原点的距离：

$$\mathcal{L}_o(\hat{z}) = -\frac{1}{\sqrt{N_z}} \|\hat{z}\|_2 \tag{3-9}$$

随后定义异常度指标为 $\mathcal{L}_{\mathrm{n}}$ 与 $\mathcal{L}_{\mathrm{o}}$ 之间的凸组合（convex combination）：

$$a\big(\mathcal{Q}, \mathcal{G}(\hat{z})\big) = \lambda \mathcal{L}_n\big(\mathcal{Q}, \mathcal{G}(\hat{z})\big) + (1-\lambda)\big(\mathcal{L}_o(\hat{z})\big) \tag{3-10}$$

式中：$\lambda \in [0,1]$。如果 $a\big(\mathcal{Q}, \mathcal{G}(\hat{z})\big) > \alpha$，样本将被归类为异常样本。

3.4 算法实现

基于 CNN 和 GAN 的工业图像异常检测算法（使用 WGAN-GP 损失）训练阶段：
$\lambda = 10,\ n_{\mathrm{critic}} = 5,\ \alpha = 0.0001,\ \beta_1 = 0,\ \beta_2 = 0.9.$

输入： 梯度惩罚项系数 λ，判别器迭代系数（单位 epoch 对应的判别器迭代次数与生成器迭代次数之比）n_{critic}，批尺寸 batch size m，Adam 优化器超参数 α, β_1, β_2.

输入： 初始生成器参数$\theta_{\mathcal{G}}$，初始判别器参数$\theta_{\mathcal{D}}$.

while $\theta_{\mathcal{G}}$ 未收敛 **do**
  for $t = 1, \ldots, n_{\mathrm{critic}}$ **do**
    for i $= 1, \ldots, m$ **do**
      采样真实数据 $x \sim p_{\mathrm{data}}$，隐向量 $z \sim p_z$，随机数 $\epsilon \sim U[0,1]$.
      $\hat{x} \leftarrow \epsilon x + (1-\epsilon)\mathcal{G}(z)$
      $\mathcal{L}^i \leftarrow \mathcal{D}_{\theta_{\mathcal{D}}}\left(\mathcal{G}_{\theta_{\mathcal{G}}}(z)\right) - \mathcal{D}_{\theta_{\mathcal{D}}}(x) + \lambda\left(\left(\left\|\nabla_{\hat{x}}\mathcal{D}_{\theta_{\mathcal{D}}}(\hat{x})\right\|_2 - 1\right)^2\right)$
    end for
    $\theta_{\mathcal{D}} \leftarrow \mathrm{Adam}\left(\nabla_{\theta_{\mathcal{D}}} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}^{(i)}, \theta_{\mathcal{D}}, \alpha, \beta_1, \beta_2\right)$

9: **end for**

10: 采样隐向量批次 $\{z^i\}_{i=1}^m \sim p_z$.

11: $\theta_{\mathcal{G}} \leftarrow \text{Adam}\left(\nabla_{\theta_{\mathcal{G}}} \frac{1}{m}\sum_{i=1}^m -\mathcal{D}_{\theta_{\mathcal{D}}}\left(\mathcal{G}_{\theta_{\mathcal{G}}}(z)\right), \theta_{\mathcal{G}}, \alpha, \beta_1, \beta_2\right)$

12: **end while**

4 实验与算法评估

本文使用了 3 个不同的数据集以对本文算法进行实验评估：MNIST 手写体数字识别数据集、CIFAR-10 自然图像数据集和 MVTecAD 工业异常检测数据集。相关评估表明，其异常分类任务的 F1 指标（平均值）分别为 98.76%、89.16% 和 79.45%，且均能够生成较高质量的重建图像和异常区域掩膜，可以同步解决对于异常图像的分类、定位和分割问题，在异常检测任务中提供一致的有效结果。

4.1 实验数据集

4.1.1 MNIST 手写体数字识别数据集

MNIST 手写体数字识别数据集[①]来自美国国家标准与技术研究所(National Institute of Standards and Technology（NIST)），由 LeCun 等人于 1998 年发布[37]，是图像分类任务的典范数据集。数据集由 250 人手写的数字构成，是 NIST 特殊数据集 3（由美国人口普查局员工书写的数字）和特殊数据集 1（由高中生书写的数字）的子集变种。

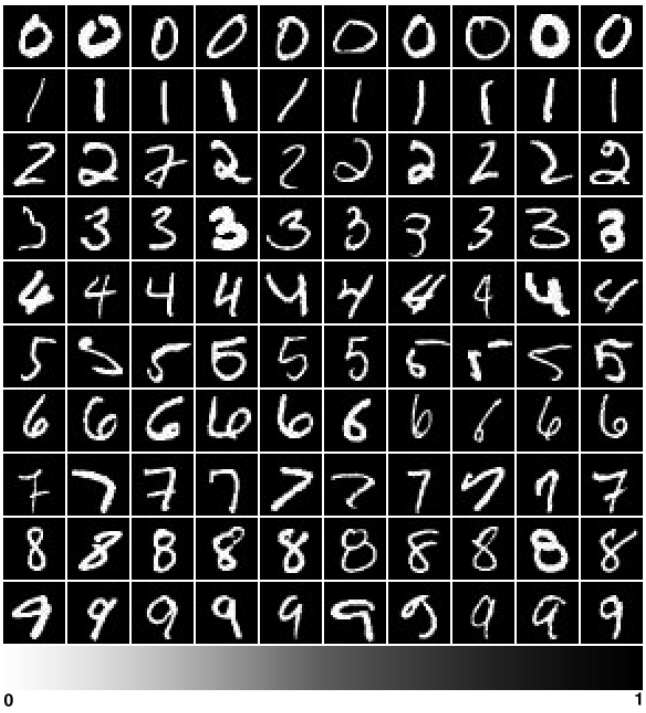


图 4-1 MNIST 手写体数字识别数据集图像示意。

① http://yann.lecun.com/exdb/mnist/

NIST 采集的原始的黑白二值手写数字图像均经过尺寸标准化至固定大小的 20×20 像素框中，同时保持其纵横比。MNIST 数据集通过计算像素的质心并平移图像以将该点定位在 28×28 尺寸图像的中心。MNIST 的标准化算法使用抗锯齿技术，因此生成的图像包含灰度级，像素深度数值均被归一化至 0－1 范围。训练集由 60000 张 28×28 尺寸的 10 类图像（对应于数字“0”－“9”，每类 6000 张图像）组成，测试集由同样比例的 10000 张测试图像（每类 1000 张图像）组成。

关于异常图像检测的部分文献使用了 MNIST 数据集作为初步的测试参考，使用某一类别的数字作为正常样本进行训练，并在测试集中混杂各类别数字以简单模拟异常的多样性。本文工作使用训练集中的各类数字分别作为正常样本进行训练；测试集组成保持不变，由 10000 张测试图像组成，以此形成了原始 MNIST 数据集的子集，以模拟真实工业环境下只使用正常样本进行学习、同时要求对没见过的异常样本具有鉴别能力的场景。

4.1.2 CIFAR-10 自然图像数据集

CIFAR-10（加拿大高等研究院（Canadian Institute for Advanced Research），10 类图像）数据集②由 Krizhevsky 等人于 2009 年发布[38]，是 Tiny Images 数据集的子集，由 60000 张 32×32 尺寸的彩色图像，其中 50000 张训练图像（每类 5000 张训练图像）和 10000 张测试图像（每类 1000 张测试图像）组成。

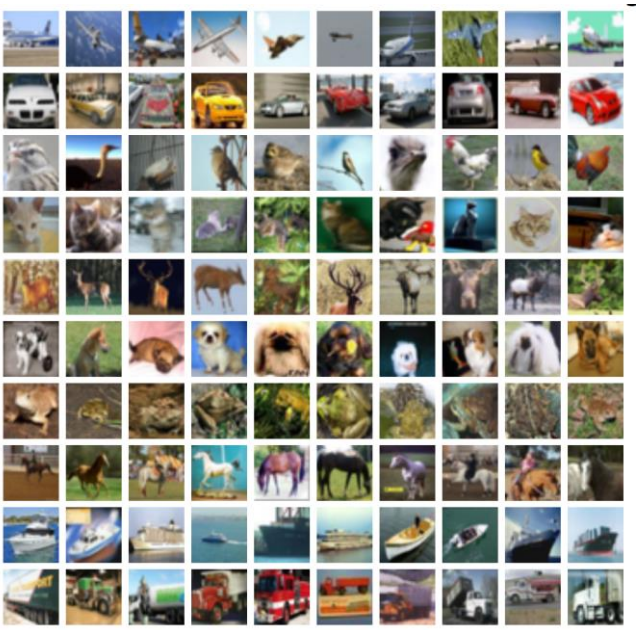


图 4-2 CIFAR-10 自然图像数据集图像示意②

② https://www.cs.toronto.edu/~kriz/cifar.html

图像来自 10 个互斥的类别：飞机、汽车（不含卡车或皮卡）、鸟、猫、鹿、狗、青蛙、马、船和卡车（不含皮卡）。前页图 4-2 给出了 CIFAR-10 自 · 然图像数据集中部分图像的示意，自上至下依序分别为前述的 10 个类别的图像。

关于异常图像检测的大量文献均使用了 CIFAR-10 数据集作为模型质量评估的基准，即使用某一类别的自然图像作为正常样本进行训练，并在测试集中混杂其他各类别自然图像，以深度模拟异常的高度复杂性和多样性。本本文工作中使用了数据集的子集：来自汽车类别的图像被视为正常样本，而来自所有其他类别的图像被视为异常样本。测试集由 1000 个正常测试样本（汽车）和 1000 个随机选择的来自其他所有类别的异常测试样本组成。

4.1.3 MVTecAD 工业异常检测数据集

MVTecAD 工业异常检测数据集③是由机器视觉软件公司 MVTec 的 Bergmann 等人于 CVPR2019 发布的一个用于异常检测的数据集[39]，包含 5354 张不同对象和纹理类别的高分辨率真彩色图像。数据集源于工业视觉中的真实产品质检场景，由 15 个类别（地毯、格栅、皮革、瓷砖、木材；瓶子、线缆、胶囊、榛子、螺母、药丸、螺丝、牙刷、晶体管以及拉链）的产品组成，可划分为质地纹理和物体结构两个大类。每个类别的产品均含一组仅含正常无缺陷图像样本的训练集，和一组同时包含有正常和异常样本的测试集。

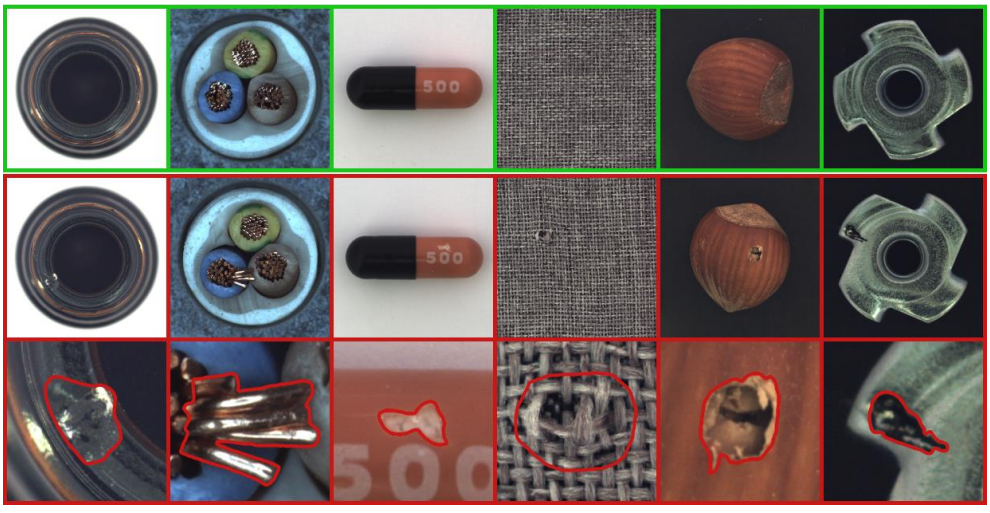


图 4-3　MVTecAD 工业异常检测数据集图像示意③

用于测试的异常图像包含有超过 70 种不同类型的缺陷，如划痕、凹痕、污染以及各种结构变化。此外，MVTecAD 数据集还为测试集中的所有异常样本提供了像素级精

③ https://www.mvtec.com/company/research/datasets/mvtec-ad/

度的真值基准（ground-truth) 区域标注。这是第一个多目标、多缺陷，同时提供像素级精度的区域标注，并专注于现实世界应用的全面的异常检测数据集。

上页图 4-3 给出了 MVTecAD 工业异常检测数据集中部分图像的示意，图中从左至右各列分别来自瓶子、线缆、胶囊、地毯、榛子和螺母类别；自上至下第 1 行为正常样本图像，第 2 行为异常样本图像，第 3 行为放大后的第 2 行对应样本的像素级精度的异常区域标注。

由于完整数据集图片分辨率过高，直接进行训练将消耗大量计算资源，因此本文对图片进行了分辨率压缩处理，使用数据集的小规模子集进行训练。

4.2 实验过程

为了评估所提出的方法，本文进行了一系列实验。在附录中给出了有关网络结构和训练配置的代码以及详细说明。在各个数据集上的实验中，所有网络都使用默认参数进行了训练，$\lambda = 10$，隐向量 z 的尺寸为 128。

在 NVIDIA GTX1080 GPU 上训练了所设计的算法模型。网络均使用了 50 的批尺寸（batch size），在 MNIST 数据集上针对不同的数字作为正常样本，各进行了 32 个时期（epoch）的训练迭代，训练时间约为 40 小时；在 CIFAR-10 数据集上进行了 50 个 epoch 的训练，训练时间约为 22 小时；在 MVTecAD 数据集上进行了 64 个 epoch 的训练，训练时间约为 36 小时。

本文算法模型默认下的实现可以接受尺寸为 28×28 的灰度图像作为输入；而 CIFAR 中的图像为三通道 30×30 尺寸的彩色图像，通过将通道数增加到 3 并分别去除生成器和判别器中的一个残差块（residual block）来调整默认的模型实现；MVTecAD 中的图像为三通道的超清彩色图像，直接在预处理阶段将训练集和测试集的所有图像压缩至 28×28 分辨率来进行实验。

4.3 实验结果与评估

4.3.1 MNIST 数据集实验结果

依次使用不同的数字作为正常样本进行了十组训练。右页图 4-4 记录了部分以数字“0”为正常样本训练的模型在测试集上由生成器重建出的图像结果和异常区域的像素级残差掩膜参考。

图中，第 1 行为输入的测试集原始图像，第 2 行为生成器生成的正常模式（“0”模式）重建图像，第 3 行为计算得出的异常区域残差掩膜参考；所有图像的像素位深度均被归一化至 0 – 1 范围。

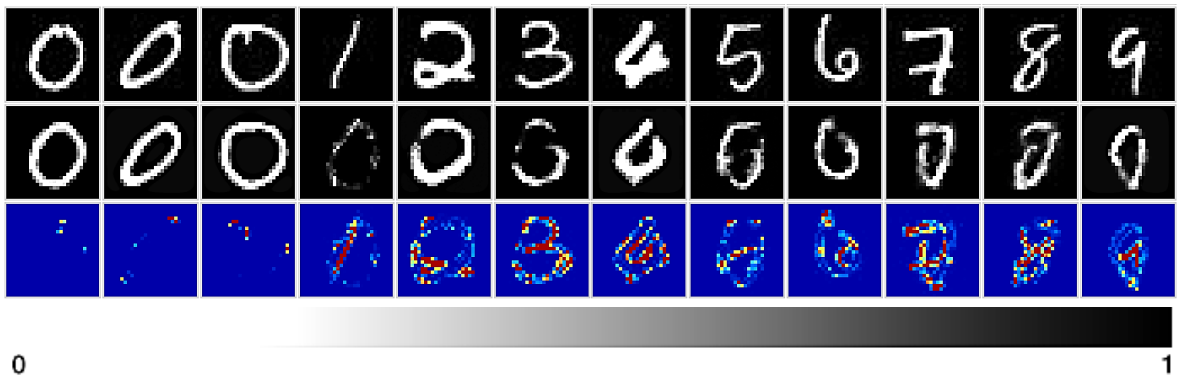


图 4-4　MNIST 数据集图像以“0”类型作为正常样本训练的生成结果

可以直观地看到，模型针对所有正常（“0”）与异常（非“0”的其他数字）的测试集样本均能重建出符合数字“0”模式的图像，并基于重建图像与输入样本图像的差异，给出了符合输入的残差掩膜结果。

这样的结果同步且高质量地完成了对于异常图像的分类、定位和分割问题：由模型根据异常分数给出的正常与异常标签的分类，以及由生成的残差图像结果给出的潜在异常区域的像素级精度的掩膜范围（如“2”、“5”、“7”的横画以及“8”的中心 x 字形交叉区域）。

表 4-1 和图 4-5 给出了 MNIST 数据集使用不同数字作为正常样本所进行的异常检测的精度（precision）、召回率（recall）和 F1 指标，其计算过程中均将异常样本视为阳性。由于模型分别只使用不同数字的训练子集上进行训练，因此不同数字下的模型学得的划分所依据的异常分数不同。

表 4-1 在 MNIST 数据集中使用不同数字作为正常样本进行异常检测的性能指标

模式	0	1	2	3	4	5	6	7	8	9	平均
精度	99.94%	**99.95%**	99.79%	99.84%	99.93%	99.89%	99.86%	99.93%	99.93%	99.83%	99.89%
召回率	**99.18%**	95.11%	97.02%	98.34%	99.09%	97.51%	98.16%	95.74%	98.73%	97.66%	97.65%
F1 指标	**99.56%**	97.47%	98.39%	99.09%	99.51%	98.68%	99.00%	97.79%	99.33%	98.73%	98.76%

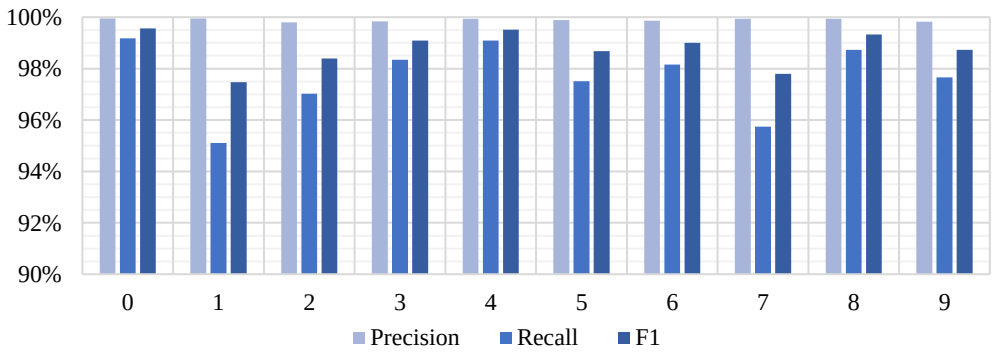


图 4-5　在 MNIST 数据集中使用不同数字作为正常样本进行异常检测的性能指标

注意到模型在所有类别中均取得了很高的精度指标，但在 1、7 二类数据的召回率表现较低，这是由于这两类图像的模式较为类似，按一方的样本进行训练的模型针对另一方进行的重建图像结果差异不大，造成该类异常数据容易漏检、误判；而真实的工业生产场景中少有出现类间差异小的异常情况，因此模型仍能保障自身鲁棒性。

4.3.2 CIFAR-10 数据集实验结果

使用汽车作为正常样本进行了训练，图 4-6 记录了模型在测试集上由生成器重建出的图像结果和异常区域的像素级残差掩膜参考。

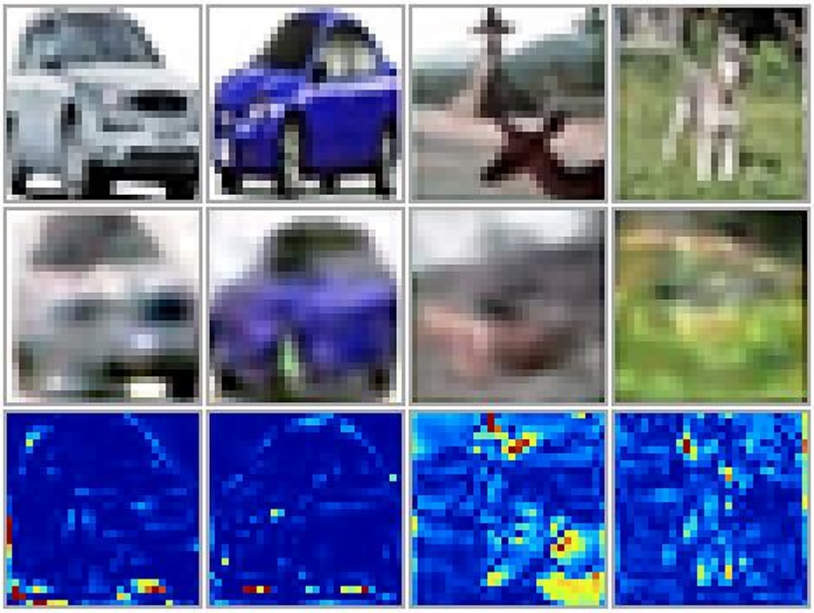


图 4-6　CIFAR-10 数据集图像以汽车类型作为正常样本训练的生成结果

图中，第 1 行为输入的测试集原始图像，第 2 行为生成器生成的正常模式（汽车模式）重建图像，第 3 行为计算得出的异常区域残差掩膜参考。

可以直观地看到，模型针对所有正常的测试集样本（从左至右第 1、2 列的汽车）均能重建出虽然模糊、但仍符合汽车模式的图像，而对于未见过的其他异常样本（第 3、4 列）则仍尝试按照汽车模式进行重建，产生无明显模式的模糊图像。基于重建图像与输入样本图像的差异，给出了符合输入的残差掩膜结果。

这样的结果依然同步且高质量地完成了对于异常图像的分类、定位和分割问题：由模型根据异常分数给出的正常与异常标签的分类，以及由生成的残差图像结果给出的潜在异常区域的像素级精度的掩膜范围。

在由 1000 个汽车和 1000 个随机选择的其他类别自然图像组成的测试集中，模型在异常检测任务上达成了 91.21% 的精度（precision）、87.20% 的召回率（recall）和 89.16% 的 F1 指标，计算过程中均将异常样本（非汽车）视为阳性。

4.3.3 MVTecAD 数据集实验结果

对于每一类别分别使用其对应的正常样本进行了训练。图 4-7 记录了模型在线缆、瓶子、榛子 3 种物体结构和地毯、皮革、木材 3 种质地纹理的类别上由生成器重建出的图像结果和异常区域的像素级残差掩膜参考。

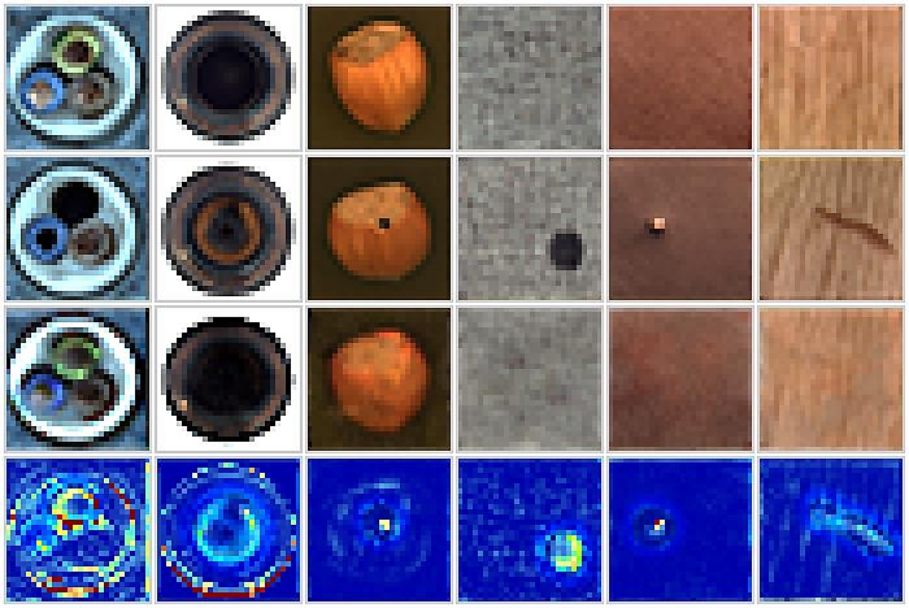


图 4-7　MVTecAD 图像以各列相应类别的正常样本训练的生成结果

图中，第 1 行为训练集正常样本数据原始图像，第 2 行为测试集异常样本原始图像，第 3 行为其对应的生成器重建出的正常模式（每列均对应于各自的物体结构和质地纹理模式）图像结果，第 4 行为计算得出的异常区域像素级残差掩膜参考。

可以直观地看到，模型针对未见过的拥有破损、缺失、刮痕等多种不同的异常样本均成功重建出了符合其相应的正常样本模式的图像。基于重建图像与输入样本图像的差异，给出了非常符合输入图像中异常位置的残差掩膜结果。

每一类别均使用了 100 张正常图像作为无异常的训练集输入样本、100 张正异常混杂样本作为测试集样本。模型在异常检测任务上达成了 81.12% 的精度（precision）、77.84% 的召回率（recall）和 79.45% 的 F1 指标，计算过程中均将异常样本视为阳性。

由于 MVTecAD 图像基本均系流水线采集的高度相似的标准化图像，其正常样本空间分布紧密，因而对于异常图像的重建和判别具有优势，可以使用较少的样本便能生成质量较高的重建图像。注意到模型在榛子以及质地纹理类别的图像中能取得很高质量的异常区域掩膜，但在线缆、瓶子等色彩及亮度变化大、形状复杂的结构化图像中受噪声影响较大，换用高质量图像进行训练或进行更高次迭代有助于改善这类异常图像的重建效果。

5 结论与展望

5.1 研究工作总结

本文综合了深度学习有关算法与图像异常检测有关技术，设计并实现了由卷积神经网络（CNN）和生成对抗网络（GAN）综合构成的深度神经网络工业异常检测算法系统主体模型。相较于单独将 CNN 或 GAN 模型引入并应用到基于机器视觉的工业异常检测场景中，该算法综合提取了两个习用模型各自的优点并加以结合，且相应地遵循了对于小样本问题和背景噪声等图像异常检测难点的解决途径，能够有效地达成高可解释性、非平衡样本适应性、高召回率和低假阳率等系统控制要素，在一定程度上解决了目前已有的部分算法的不足之处。

在 MNIST 手写体数字识别、CIFAR-10 自然图像以及 MVTecAD 工业异常检测数据集上对本文的有关算法进行的部署、测试和结果。相关评估表明，其异常分类任务的 F1 指标（平均值）分别为 98.76%、89.16% 和 79.45%，且均能够生成较高质量的重建图像和异常区域掩膜，可以同步解决对于异常图像的分类、定位和分割问题，在异常检测任务中提供一致的有效结果。

5.2 研究工作展望

在学界和业界的共同努力下，工业图像异常检测领域已有许多行之有效的算法模型被提出并部署至相关领域的实际生产过程中，基于深度学习的异常检测方法更是其中的热点；但囿于实际生产过程的复杂性，依然有大量问题亟待解决。本节结合基于深度学习的异常检测方法研究现状，联系实际结果，在空间尺度上对算法在相关领域的应用价值和在实际中推广部署可能性进行了横向展望；同时在时间尺度上对未来进行深入研究的方向及可能遇到的难点进行纵向展望。

5.2.1 相关应用前景

工业异常检测应用作为机器视觉的重要组成部分，其研究焦点已逐渐从经典的图像处理和机器学习方法过渡转移到深度学习方法，将过去分多个步骤和环节零散处理的分类、定位和分割任务统一为一个算法模型。这样的趋势已经逐步走进成熟的工业应用，如流水线的异常检测所处理的产品大多高度标准化，使用简单的基于深度学习的异常检测算法即可达成，解决了过去传统方法往往无法解决的大量难题，预期可以使得工业异常检测更加便捷、智能，提高相关部门的生产效益、效率和行业竞争力。

在未来，类似的算法可以部署到电缆、铁轨、墙体以及路面等基础设施的大规模维护和检查，以及处理器芯片与印刷电路板等高度复杂且精密的元器件的检查和清理工作之中，发展前景广阔。

5.2.2 未来工作展望

1、针对样本均衡问题继续对有关模型做出改进。使用基于特征感知的分类算法需要大量且平衡的正负样本，而使用基于图像生成与重建的算法则完全不需要负样本。在真实工业场景中，可用的训练样本情况往往介于二者之间：有大量的正常样本，同时少量的异常样本则在初期难以被大量提供、且很难搜集到所有类型的异常样本，但可以在检测过程中逐步积累，补足数据，进行类似于自监督的学习过程。现有模型多难以与此场景兼容，因此还有待学界进一步参考仿人视觉认知系统与类脑计算的等先验知识的引入，指导异常检测网络模型的训练和学习，提出更加贴切于工业图像异常检测真实需求的算法模型。

2、网络结构参数调整。目前已有的算法在部署过程中多需要手动进行网络结构的改动和具体参数的调整，耗费精力且难以逼近最优解。自动机器学习和网络架构搜索技术的发展可以使得机器搜寻和自动生成的网络逐步替代人工设计的结构参数，能够进一步解放生产力、提升模型部署的效率和结果的性能表现。

3、对异常进行分类。在真实的工业场景中，往往需要对不同的异常进行分类，将现有的异常-正常二分类问题扩展成了多分类问题。此外，随着检测过程的不断进行，可能有先前未见过的异常类型出现，此时依然要求模型能够保障给出正确的判断，并在新类型的异常被即时地归类并补充进系统后继续保持这一能力。这要求算法同时对特定和宽泛的的样本类型均具有响应能力。

4、对高分辨率图像进行高质量的重建和掩膜生成。一些工业场景需要对大尺寸高分辨率图像进行异常定位和分割，要求基于生成的算法能够适应对大尺度图像进行高质量的重建图像生成工作，并给出形状复杂或者多个分散的异常定位覆盖区域。目前对于高质量图像的生成依赖于大量训练数据和相当高的训练成本，且因为每次检测都需要重新生成图像，故图像生成过程本身亦可能会耗费不可忽视的成本。

5、联邦学习。工业异常检测任务往往缺乏大量优质可用的数据集，与此同时在真实场景中可用的图像样本数量相当庞大且具有共通特性，例如污损异常广泛存在于各种光滑表面的工业生产过程中。目前的这部分数据没有得到有效的利用，而基于异域数据的联邦学习能够打破不同工业异常检测应用场景之间的壁垒，充分利用已有资源学习可用的数据来提升网络性能，为一未来的潜在发展方向。

6 致　谢

首先由衷感谢软件所毕业设计工作组的指导老师们，他们在项目设计、中期答辩乃至论文草稿和定稿的审核过程中尽心尽力、付出良多，这篇毕业设计论文之所以得以完成，离不开老师们悉心指导和耐心督促的功劳。令我记忆犹新的是，草稿审核的过程中，老师在强调参考文献时所说的那句“要把论文写在祖国大地上”，让我收获了求学之路和人生之路的启示，真切地体会到了“师之大者，为国为民”的品格风范，深受感动和激励。

感谢指导师兄师姐们对于本文初期任务规划以及试验工作提供的支持和帮助，在一同商讨模型细节、谋篇布局、遣词用句的过程中倾注了大量的热情和心血，使我学习到了不少知识和技能，感触良多，真正体会到了“幸福都是奋斗出来的”。

感谢在西安交通大学本科求学期间的遇到的诸位老师在五年期间对我的谆谆教导。感谢诸位同学们在少年班和计算机系的学习过程中的相互鼓励、相互帮助。感谢父母一直以来对我的无条件支持与爱。

由衷地感谢所有在西安交通大学学习期间和本文研究与写作过程中给予帮助的人。

风雨流年，匆匆而逝。祝愿西安交通大学的明天更加美好！

7 参考文献

[1] Chalapathy R, Chawla S. Deep Learning for Anomaly Detection: A Survey[J]. 2019.

[2] VARUN, CHANDOLA, ARINDAM, et al. Anomaly Detection: A Survey[J]. Acm Computing Surveys, 2009.

[3] 张笑璐,邹益胜,曾大懿,彭飞,赵市教.样本不均衡下的 DCGAN 轴承故障诊断方法[J/OL].机械科学与技术:1-8.https://doi.org/10.13433/j.cnki.1003-8728.20200335.

[4] 刘定祥,乔少杰,张永清,韩楠,魏军林,张榕珂,黄萍.不平衡分类的数据采样方法综述[J].重庆理工大学学报(自然科学),2019,33(07):102-112.

[5] 周建伟. 不平衡数据的下采样方法研究[J]. 计算机与数字工程, 2019, 47(9):2155-2160.

[6] 石洪波,陈雨文,陈鑫.SMOTE 过采样及其改进算法研究综述[J].智能系统学报,2019,14(06):1073-1083.

[7] Oksuz K, Cam B C, Kalkan S, et al. Imbalance Problems in Object Detection: A Review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP(99):1-1.

[8] Tao X, Wang Z, Zhang Z, et al. Wire Defect Recognition of Spring-Wire Socket Using Multitask Convolutional Neural Networks[J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2018, 8(4):689-698.

[9] Liu L, Cao D, Wu Y, et al. Defective samples simulation through adversarial training for automatic surface inspection[J]. Neurocomputing, 2019, 360(Sep.30):230-245.

[10] Ruoxu, Ren, Terence, et al. A Generic Deep-Learning-Based Approach for Automated Surface Inspection.[J]. IEEE transactions on cybernetics, 2017.

[11] Kim S, Kim W, Noh Y K, et al. Transfer learning for automated optical inspection[C]// International Joint Conference on Neural Networks. IEEE, 2017.

[12] Tabernik D, Ela S, Skvar J, et al. Segmentation-based deep-learning approach for surface-defect detection[J]. Journal of Intelligent Manufacturing, 2020, 31.

[13] 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.

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[15] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[16] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. IEEE Computer Society, 2014.

[17] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[18] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// Computer Vision & Pattern Recognition. IEEE, 2016.

[19] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[20] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[21] He K, Gkioxari G, P Dollár, et al. Mask R-CNN[C]// IEEE Transactions on Pattern Analysis & Machine Intelligence. IEEE, 2017.

[22] ROOPAK, SHAH, EDUARD, et al. SIGNATURE VERIFICATION USING A "SIAMESE" TIME DELAY NEURAL NETWORK[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1993, 07(4):669-669.
[23] Berg A, Felsberg M, Ahlberg J. Unsupervised adversarial learning of anomaly detection in the wild[J]. Proceedings of the Frontiers in Artificial Intelligence and Applications, Santiago de Compostela, Spain, 2020, 325: 1002-1008.
[24] Lu Y, Xu P. Anomaly Detection for Skin Disease Images Using Variational Autoencoder[J]. 2018.
[25] Mei S, Yang H, Yin Z. An Unsupervised-Learning-Based Approach for Automated Defect Inspection on Textured Surfaces[J]. IEEE Transactions on Instrumentation and Measurement, 2018:1266-1277.
[26] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[C]// Computer ence. 2015.
[27] Yan X, Cao X, Fang W, et al. Learning Discriminative Reconstructions for Unsupervised Outlier Removal[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.
[28] An J, Cho S. Variational Autoencoder based Anomaly Detection using Reconstruction Probability.
[29] Akcay S, Atapour-Abarghouei A, Breckon T P. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training[M]. 2019.
[30] Niu S, H Lin, Niu T, et al. DefectGAN: Weakly-Supervised Defect Detection using Generative Adversarial Network[C]// 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE). IEEE, 2019.
[31] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. MIT Press, 2014.
[32] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[J]. 2017.
[33] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. arXiv preprint arXiv:1704.00028, 2017.
[34] Mao X, Li Q, Xie H, et al. Least Squares Generative Adversarial Networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.
[35] Berthelot D, Schumm T, Metz L. BEGAN: Boundary Equilibrium Generative Adversarial Networks[J]. arXiv, 2017.
[36] Lucic M, Kurach K, Michalski M, et al. Are GANs Created Equal? A Large-Scale Study[J]. 2017.
[37] Lecun Y, Bottou L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[38] Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. 2012.
[39] Bergmann P, Fauser M, Sattlegger D, et al. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[40] 陈亮，吴攀，刘韵婷,等. 生成对抗网络 GAN 的发展与最新应用[J]. 电子测量与仪器学报，2020(6):70-78.

8 附 录

附录 A 外文翻译原文及其译文

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Unsupervised Adversarial Learning of Anomaly Detection in the Wild

Amanda Berg[1,2] and **Michael Felsberg**[2] and **Jörgen Ahlberg**[1,2]

Abstract. Unsupervised learning of anomaly detection in high-dimensional data, such as images, is a challenging problem recently subject to intense research. Through careful modelling of the data distribution of normal samples, it is possible to detect deviant samples, so called anomalies. Generative Adversarial Networks (GANs) can model the highly complex, high-dimensional data distribution of normal image samples, and have shown to be a suitable approach to the problem. Previously published GAN-based anomaly detection methods often assume that anomaly-free data is available for training. However, this assumption is not valid in most real-life scenarios, a.k.a. in the wild. In this work, we evaluate the effects of anomaly contaminations in the training data on state-of-the-art GAN-based anomaly detection methods. As expected, detection performance deteriorates. To address this performance drop, we propose to add an additional encoder network already at training time and show that joint generator-encoder training stratifies the latent space, mitigating the problem with contaminated data. We show experimentally that the norm of a query image in this stratified latent space becomes a highly significant cue to discriminate anomalies from normal data. The proposed method achieves state-of-the-art performance on CIFAR-10 as well as on a large, previously untested dataset with cell images.

1 Introduction

Anomaly detection is the identification of *rare* samples, objects, or events that are regarded as anomalous compared to what is considered to be normal. Anomalies are sometimes also referred to as outliers [21]. Due to the quite general problem formulation, anomaly detection is applicable to a wide range of different fields, such as e.g. agriculture [10], medicine [33, 32], and finance [1, 2]. In the context of machine learning, anomaly detection can be *supervised*, *semi-supervised*, or *unsupervised*. This paper addresses *unsupervised* anomaly detection.

The objective of unsupervised anomaly detection is to detect previously unseen rare objects or events without any prior knowledge about these. The only information available is that the percentage of anomalies in the dataset is small, usually less than 1%. Since anomalies are rare and unknown to the user at training time, anomaly detection in most cases boils down to the problem of modelling the normal data distribution and defining a measurement in this space in order to classify samples as anomalous or normal. In high-dimensional data such as images, distances in the original space quickly lose descriptive power (curse of dimensionality) and a mapping to some more suitable space is required. Due to their latent space, Generative Adversarial Networks (GANs) [19] can model complex, high-dimensional data distributions [11] and are, therefore, suitable for anomaly detection in images. GAN-based methods also provide the ability to localize anomalies within images in contrast to many classical anomaly detection methods [32, 33]. Although partly addressed in recent works [3, 4, 12, 28, 32, 33, 37, 38], unsupervised anomaly detection still remains a challenging problem.

The main limitation of these previously published unsupervised GAN-based methods is their assumption that anomaly-free data is available for training. For this reason, we argue that they are not *truly* unsupervised, since completely anomaly-free data requires weak labelling. Anomaly contamination of GAN training data is expected to reduce detection performance [7]. In this work, we show that this is indeed the case for a recent, state-of-the-art GAN based anomaly detection method f-AnoGAN [32] and its variations.

Further, we show using t-SNE visualization [35] that anomalous and normal validation samples are scattered in latent space such that the GANs expressiveness with respect to classification is limited. To mitigate this problem, an image-to-latent-space encoder trained *jointly* with the generator is proposed. The joint training coupled with an image distance encoder loss enforces similar images to lie close to each other also in latent space. In this stratified latent space, latent vectors of anomalous samples prove to have shorter norms than those of normal samples. We show this empirically in a number of experiments on two datasets, based on CIFAR-10 and on a large cell-image dataset. Our approach achieves state-of-the-art performance in both cases.

Contributions

- We conduct an empirical study varying the amount of anomalies in the training data and measure the degradation of the anomaly detection in existing methods.
- We propose an approach to truly unsupervised anomaly detection based on simultaneous encoder training that improves results even when the training data is contaminated with anomalies.

2 Related work

Anomaly detection is an important problem relevant to a vast number of fields, e.g. malware intrusion detection [24], retinal damage detection [32, 33], and detection of anomalous events in surveillance videos [34]. A complete review of anomaly detection methods is beyond the scope of this paper, the interested reader is referred to [8, 9]. In the particular case of *unsupervised* anomaly detection, labels are unknown at training time. This paper is focused on unsupervised deep learning based anomaly detection of/in high-dimensional, non-sequential data with spatial coherence, i.e., images.

[1] Termisk Systemteknik AB, Sweden
Email: {amanda.jorgen.ahl}berg@termisk.se
[2] Computer Vision Laboratory, Linköping University, Sweden
Email: {amanda.jorgen.ahl,michael.fels}berg@liu.se

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Classical methods for unsupervised anomaly detection include probabilistic methods that model the data distribution, e.g., by using a non-parametric Kernel Density Estimator (KDE) [29] as in [13] where it is applied to intrusion detection. Samples in low density areas are treated as anomalies. Another example of a probabilistic, parametric method is the RX anomaly detector [31]. Due to the *curse of dimensionality*, probabilistic methods are, however, not suitable for high-dimensional data such as images. Also, they typically do not provide the ability to *localize* anomalies in images.

In contrast, reconstruction-based methods provide the possibility to localize anomalies within images. The aim of these methods is to find a lower-dimensional latent space from which normal samples can be reconstructed. A query image is then projected onto this latent space and the reconstructed image is compared to the query image by some image distance measurement in order to discriminate anomalous cases. The latent space can be modelled using, e.g., Auto Encoders [36], Variational Auto Encoders [5], or Generative Adversarial Networks (GANs) [4, 12, 32, 33, 37, 38]. In the context of unsupervised anomaly detection, GANs were first introduced by Schlegl et. al. [33] (AnoGAN). They proposed to use a combination of the l_2-norm and a discrimination loss between a query image and its closest reconstruction match as an anomaly score. Based on this approach, Deecke et. al. [12] proposed a similar method (ADGAN) that improved the results slightly. In contrast to AnoGAN, ADGAN initialized the search in latent space for the closest match at multiple locations. Recently, and concurrent to this work, Schlegl et. al. [32] proposed f-AnoGAN, improving their method (AnoGAN) by replacing the Deep Convolutional GAN (DCGAN) [30] with a Wasserstein GAN (WGAN-GP) [20] and they also introduced an encoder that was trained separately for image to latent space mapping. The usage of an encoder instead of an iterative optimization procedure in order to speed up image to latent space mapping has also been explored by Zenati et. al. [37, 38] who employed an architecture similar to a Bidirectional GAN (BiGAN) [14] with pairs of (X, z) as input to the discriminator. We argue that the novelty of the proposed method compared to [37, 38] is the discussion of the impact of such an encoder on the structure of the latent space, and also the problem of training data contamination.

Ngo et al. [28] make the observation that the usual GAN objective encourages the distribution of generated samples to overlap with the real data, which may not be optimal in the case of anomaly detection. They further propose an *encirclement* loss that places generated samples at the boundary of the distribution and can then use the discriminator directly to discriminate anomalous samples.

Golan and El-Yaniv [18] proposed another type of method trained to map input images to a set of geometric transformations. In contrast to the reconstruction-based methods, it can not provide anomaly localization in images.

Some of the methods mentioned above [4, 12, 32, 33] claim to be unsupervised while at the same time assuming anomaly-free data for training. The acquisition of anomaly-free data requires labelling of data as normal. However, anomalous objects and/or events are rare and difficult to label in most real-world scenarios.

Beggel et. al. [7] conclude that the anomaly detection performance is reduced when the training set is contaminated with anomalies. They use an Adversarial Auto Encoder [26] to mitigate the problem by rejecting potential anomalies already during training. The proposed method improves detection results in the case of anomalies present in the training data in a different way. Instead of rejecting, we propose to use an encoder trained jointly with the GAN. As we show in our experiments, the anomalies need not to be rejected at training time, but mapped closer to the origin.

3 Method

The architecture of the proposed method is a combination of the progressive growing GAN (pGAN) [22] and ClusterGAN [27] but without class labels. An overview of the architecture at both training and testing time is presented in Figure 1. The generator and discriminator are equal to the ones in pGAN [22], while the encoder was inspired by ClusterGAN [27]. The architecture and objective function is further described below. At test time, the discriminator is discarded and the parameters of the generator and encoder are fixed. A query image Q is considered to be anomalous or not based on an anomaly score a.

3.1 Network architecture

One of the major drawbacks of AnoGAN [33] is its reliance on accurate reconstruction by a DCGAN [30]. DCGANs are, among other things, known to suffer from mode collapse [6]. For that reason, the inventors of AnoGAN replaced the DCGAN with a WGAN-GP [20] in f-AnoGAN. We instead propose to employ a progressive growing GAN (pGAN) [22]. pGAN also employs the WGAN-GP loss but incrementally adds new layers to the generator and discriminator while training. This approach has proven to increase the stability and robustness of a GAN, especially in the case of high-resolution images. The generator $\mathcal{G}(z : \theta_G)$ $\mathcal{G} : z \mapsto X_g$ and discriminator $\mathcal{D}(X : \theta_D)$ $\mathcal{D} : X \mapsto Y$ of the proposed method are equal to the ones used in pGAN. The prior, $z \sim \mathcal{N}(0, \mathrm{I}) \in \mathbb{R}^{N_z}$ is drawn from a Gaussian distribution.

Another update in f-AnoGAN compared to AnoGAN was the introduction of an encoder instead of the iterative search, which greatly improved detection speed. The encoder $\mathcal{E}(X : \theta_E)$ maps images to latent space $\mathcal{E} : X \mapsto \hat{z}$. In contrast to f-AnoGAN, the proposed method suggests to train the encoder $\mathcal{E}$ together with $\mathcal{G}$ and $\mathcal{D}$ in the same progressive manner as $\mathcal{G}$ and θ_G and θ_E are updated jointly. Various training strategies to learn an encoder have been explored by Dumoulin et. al. [15], although on different problems, and they emphasized the importance of learning $\mathcal{G}$ and $\mathcal{E}$ jointly. We make the same observation in our experiments.

Deecke et. al. [12] concluded that the discriminator is unsuitable for anomaly detection. While trained to separate real from generated images, thus forcing the two probability distributions to overlap, it is not trained to handle anomalous samples drawn from a different distribution. At test time, see Figure 1b, $\mathcal{D}$ is discarded and the parameters of $\mathcal{G}$ and $\mathcal{E}$, θ_G and θ_E, are fixed.

3.2 Objective function

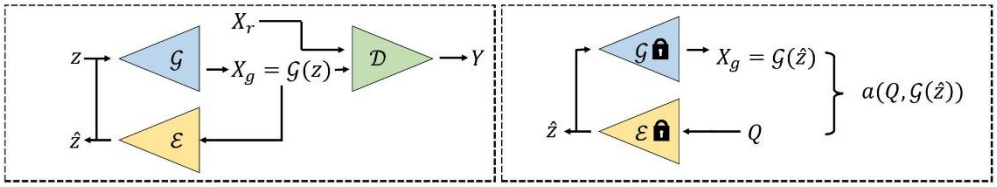
Similar to f-AnoGAN and pGAN, we employ the WGAN-GP loss [20]. However, $\mathcal{E}$ is trained jointly with $\mathcal{G}$, not in a subsequent step as in f-AnoGAN. The GAN objective for the proposed method takes the following form:

$$\min_{\theta_G, \theta_E} \max_{\theta_D} \mathop{\mathbb{E}}_{X \sim p_{\text{data}}} q(\mathcal{D}(X)) + \mathop{\mathbb{E}}_{z \sim p_z} q(1 - \mathcal{D}(\mathcal{G}(z))) + \mathop{\mathbb{E}}_{z \sim p_z} \left\| (\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z)))) \right\|_1 \tag{1}$$

where $q(x) = x$ since we use a Wasserstein loss [27]. The third term, $\mathop{\mathbb{E}}_{z \sim p_z} \left\| \mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z))) \right\|_1$, is new compared to previous works [20, 22, 32].

In contrast to BiGAN and ALI [15], the proposed architecture allows $\mathcal{G}$ and $\mathcal{E}$ to interact with each other during training, similar to the

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain



(a) Training (b) Testing

Figure 1: An overview of the proposed architecture at (a) training and (b) testing time. The encoder $\mathcal{E}$ is trained jointly with the generator $\mathcal{G}$. At test time, the discriminator $\mathcal{D}$ is discarded and the parameters of $\mathcal{G}$ and $\mathcal{E}$ are fixed. A query image Q is encoded and compared to its reconstruction $\mathcal{G}(\mathcal{E}(Q))$ in order to find an anomaly score a.

encoder used in ClusterGAN. However, while ClusterGAN computes the encoder loss in the latent space $z - \mathcal{E}(\mathcal{G}(z))$, we instead choose to compute the encoder loss in image space $\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z)))$. The by $\mathcal{G}$ reconstructed query image Q should be the closest match in image space to Q rather than the closest match in latent space, since the anomaly score a, see next section, is partly based on a distance measure in image space. Also, the image space loss structures the latent space in a different way than the latent space loss, separating normal and anomalous samples, see the evaluation section.

3.3 Anomaly detection

We propose to use an anomaly score consisting of two terms, a normalized *residual* and an *origin distance* loss. The *residual* loss $\mathcal{L}_n$ for the query image $Q \in [0,1]^{W \times H \times D}$ is defined as the ℓ_2-norm between Q and its closest match $\mathcal{G}(\hat{z})$:

$$\mathcal{L}_n(Q, \mathcal{G}(\hat{z})) = \frac{1}{N_X} \|w(Q) - w(\mathcal{G}(\hat{z}))\|_2 \tag{2}$$

where $\hat{z} = \mathcal{E}(Q)$ is the encoded latent vector for image Q. In order to minimize the impact of the image contrast to the residual loss, we, unlike f-AnoGAN, propose to apply a minmax normalization $w(x)$ of images. The normalization $w(X) : [\min(X), \max(X)]^{W \times H \times D} \mapsto [0,1]^{W \times H \times D}$ where $\min(X)$ and $\max(X)$ finds the minimum and maximum elements of X, is defined as

$$w(X) = \frac{X - \min(X)}{\max(X) - \min(X)}, \tag{3}$$

where the division is element-wise and $N_X = W \cdot H \cdot D$. Without minmax normalization, low contrast samples yield low residual losses and vice versa.

Based on our observations regarding joint encoder and generator training and how that affects the structure of the latent space, we define an *origin distance* loss $\mathcal{L}_o$ as the distance in latent space from encoded vector $\hat{z}$ to the origin:

$$\mathcal{L}_o(\hat{z}) = -\frac{1}{\sqrt{N_z}} \|\hat{z}\|_2 . \tag{4}$$

The anomaly score is then defined as the convex combination between $\mathcal{L}_n$ and $\mathcal{L}_o$ as

$$a(Q, G(\hat{z})) = \lambda \mathcal{L}_n(Q, G(\hat{z})) + (1 - \lambda)(\mathcal{L}_o(\hat{z})), \tag{5}$$

where $\lambda \in [0,1]$. Samples are classified as anomalies if $a(Q, G(\hat{z})) > \alpha$.

In [32], f-AnoGAN used a convex combination of a residual loss and a *discrimination* loss as anomaly score. The discrimination loss depends on the difference between the discriminator output and the average discriminator output. In our experiments, adding the discriminator loss did not improve detection results.

4 Evaluation and results

4.1 Datasets

Two different datasets were used for evaluation in this work. The fully annotated KTH-Cellvideos dataset [17, 25], depicting different cells, and the CIFAR-10 dataset [23]. All training images were normalized to lie within range $[-1, 1]$.

4.1.1 CIFAR-10

The CIFAR-10 dataset [23] consists of 50000 $32 \times 32 \times 3$ training images in 10 classes (5000 images per class) and 10000 test images (1000 images per class). In this work, a subset of the dataset, denoted as $\text{CIFAR}_{\text{CAR}}$, was used. Images from the car class were treated as normal samples and images from all other classes as anomalous samples. The test set consisted of the 1000 normal test samples (car) and 1000 randomly chosen anomalous test samples from all other classes.

4.1.2 KTH-Cellvideos

The KTH-Cellvideos dataset [17, 25] consists of grayscale medical images featuring living cells in microscopy image sequences. About 50% of the labelled objects in the dataset is debris, e.g. bubbles, and they are labelled as such. Events such as mitosis (cell division) and apoptosis (cell death) are also labelled and segmentation masks are available for all cells. In this work, debris is treated as anomalies and cells as normal samples.

The labelled objects in the dataset were split into a training and a test set. All labelled objects (normal/debris) were cropped in a 64 by 64 neighbourhood. In addition, training samples were rotated three times by randomly generated angles. That is, each labelled object (except for the ones reserved for the test set) in the original dataset gave rise to four samples in the training dataset. In total, there were $N = N_n + N_a$ training patches where $N_n = 525657$ is the number of normal training patches and $N_a = \frac{\gamma N_n}{1-\gamma}$ the number of anomalous training patches. $\gamma \in [0,1]$ is the user-defined percentage of anomalies in the training data. The test set consisted of 256 normal test images and 256 anomaly test images.

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

4.2 Experiments

In order to evaluate the proposed method, a series of experiments was conducted. Code and detailed descriptions of network architectures and training configurations are available at `https://github.com/amandaberg/GANanomalyDetection`. For all experiments, $N_z = 512$ and $\lambda = 0.05$. Training of the proposed method was performed on an NVIDIA GTX1080 GPU, the batch size started at 128 and ended at 32 for KTH-Cellvideos and 64 for CIFAR-10. KTH-Cellvideos networks were trained for 48 epochs (6 epochs on full resolution) and CIFAR-10 networks were trained for 32 epochs (4 epochs on full resolution). Training time was about 36 hours for KTH-Cellvideos and about 12 hours for CIFAR-10.

All f-AnoGAN networks were trained with default parameters, batch size 16 and the dimension of z was 128. The KTH-Cellvideos networks were trained for 7 epochs. The training time was about 16 hours for the generator and about 1 hour for the encoder.

The default implementation of f-AnoGAN accepts images of dimension $64 \times 64 \times 1$ as input. Images in $\mathrm{CIFAR}_{\mathrm{CAR}}$ have dimension $32 \times 32 \times 3$. The default implementation was adapted by increasing the number of channels to 3 and removing one residual block in the discriminator, generator, and encoder respectively.

For dataset $\mathrm{CIFAR}_{\mathrm{CAR}}$, the f-AnoGAN generator was not able to generate visually pleasing images after seven epochs due to the low number of training samples (5000). Even training the network for as much as 70 epochs did not improve the detection performance. Therefore, since more iterations did not improve detection performance, f-AnoGAN was only trained for seven epochs for $\mathrm{CIFAR}_{\mathrm{CAR}}$.

Anomaly detection results are measured as the Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) [16].

4.2.1 *Encoder*

Training jointly vs. training separately In the AnoGAN (note: not f-AnoGAN) paper [33], an iterative search was used to find the closest match to the query image Q in latent space. The drawbacks with this approach are that a) the optimization can get stuck in local minima, and b) evaluation was time-consuming. Here, we show that when training our method without an encoder and using an iterative search similar to the one in [33], encoded validation samples lie scattered all over the latent space, see Figure 2a. There is no separation between normal and anomalous samples.

In contrast, the introduction of an encoder stratifies the latent space. For f-AnoGAN, where the encoder is trained separately, the separation of samples (according to t-SNE) appears to be somewhat worse, Figure 2b, than for the proposed method, Figure 2c. AUC scores confirming this for the two methods are presented in the anomaly score section below. We believe that the joint encoder training enforces similar images to lie close to each other also in latent space. For the t-SNE plots, a perplexity value of 30 was used and the visualizations were consistent across multiple runs.

In Figure 3, the histograms of the coefficients of the encoded latent vectors for the validation samples from the KTH-Cellvideo dataset can be found. The networks were trained with 0% anomalies in the training data. It is clear that the proposed joint encoder training spreads the coefficients more evenly across the latent space, Figure 3c. These plots also explain why the norm of the latent vector, or the distance to origin, is not a discriminative loss in the case of f-AnoGAN. For f-AnoGAN, the samples end up on a hypercube, Figure 3a-b. In contrast, the density of coefficients is higher for anomalies close to the origin for the proposed method, Figure 3c.

In what follows, we give a possible explanation why the norms of latent variables representing anomalies are empirically smaller than those of normal images. Recall $z \sim \mathcal{N}(0, \mathrm{I}) \in \mathbb{R}^{N_z}$. In the implementation of pGAN, the prior z is normalised to unit length before being processed. A normalized *random* vector $z \in \mathbb{R}^{N_z}$ drawn from $\mathcal{N}(0, \mathrm{I})$ will have small coefficients. GAN training moves data clusters in the latent space away from the origin, otherwise the discriminator would not be able to separate them from the prior distribution, i.e. the noise. The encoder maps normal samples to clusters. Assuming high intra-class variability among anomalies, anomalies will be mapped away from the clusters and end up closer to the origin, i.e. the noise, and thus have smaller coefficients similar to a *random* vector.

When the training data is contaminated with anomalies, see Figure 2d and 2e, the confusion between normal and anomalous samples increases for f-AnoGAN. This is also confirmed in Table 2, (method d) where the norm-based loss $\mathcal{L}_o$ decreases AUC for f-AnoGAN. In contrast, the proposed method maintains the separability between samples (Figure 2f) even though the training data is contaminated with as much as 2% anomalies (method h).

Distance in image space vs. distance in latent space The proposed loss for the encoder is the third term in (1), hereby denoted by d_I:

$$d_I = \left\| \mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z))) \right\|_1 . \tag{6}$$

Generated images $\mathcal{G}(z)$ are compared with their reconstructed images $\mathcal{G}(\mathcal{E}(\mathcal{G}(z))))$ in image space. Another option would be to compare the distance between the latent vector z and the reconstructed latent vector $\hat{z} = \mathcal{E}(\mathcal{G}(z))$ in the latent space:

$$d_z = \left\| z - \mathcal{E}(\mathcal{G}(z)) \right\|_1 . \tag{7}$$

Results for the proposed method using d_I and d_z are provided in Table 1 and t-SNE visualizations [35] of latent space projections are shown in Figure 4. The network was trained on the KTH-Cellvideos dataset with 0% anomalies in the training data. Comparing the distance in image space (d_I) is clearly preferable when it comes to separation of the validation samples in latent space. A good d_I implies a good d_z but the opposite is not true. We believe this is because d_I enforces similar images (in image space) to lie close to each other also in latent space. Small variations in z and $\hat{z}$ during reconstruction are forced to yield similar images.

Table 1: AUC results for the proposed method with different encoder losses, d_z and the proposed d_I, for the KTH-Cellvideos dataset.

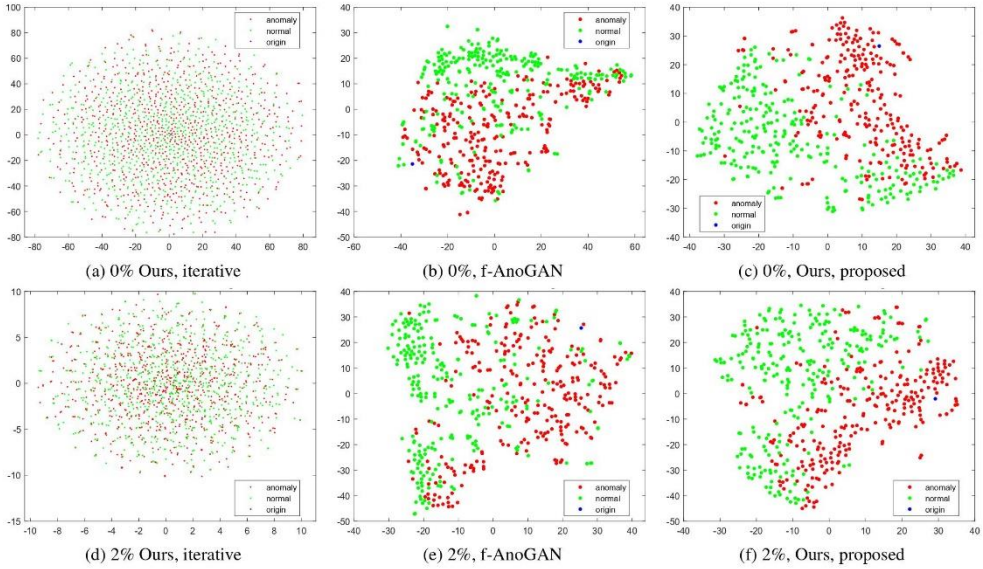
Encoder loss	$\mathcal{L}_n$	$\mathcal{L}_o$	$\mathcal{L}_n + \mathcal{L}_o$
d_I (proposed)	0.78	0.89	**0.90**
d_z	0.66	0.69	0.66

4.2.2 *Anomaly score*

As previously described, we propose to use a convex combination of a normalized residual loss $\mathcal{L}_n$ and a norm-based loss $\mathcal{L}_o$. In Table 2, AUC results for different combinations of these losses for both f-AnoGAN and our method can be seen. The networks were trained on two different datasets with two different percentages of anomalies in the training data. A is the anomaly score proposed in [32] and $\mathcal{L}_r$ is the residual loss, also from [32], without the proposed minmax normalization. Hence,

$$\mathcal{L}_r(Q, \mathcal{G}(\hat{z})) = \left\| Q - \mathcal{G}(\hat{z}) \right\|_2 . \tag{8}$$

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain



(a) 0% Ours, iterative (b) 0%, f-AnoGAN (c) 0%, Ours, proposed

(d) 2% Ours, iterative (e) 2%, f-AnoGAN (f) 2%, Ours, proposed

Figure 2: t-SNE visualization of validation samples projected to latent space for our method trained (a,d) without an encoder and iterative search for closest match, (c,f) with an encoder with latent space projection to find the closest match, and for (b,e) f-AnoGAN. The networks were tained on KTH-Cellvideos with (a-c) 0% and (d-f) 2% anomalies in the training data.

Table 2: AUC results for different anomaly losses for the proposed method and f-AnoGAN trained on three different datasets with 0% and 2% anomalies.

	Method	$\text{CIFAR}_{\text{CAR}}$ 0%	$\text{CIFAR}_{\text{CAR}}$ 2%	KTH-Cellvideos 0%	KTH-Cellvideos 2%
a)	f-AnoGAN A	0.45	0.44	0.45	0.43
b)	f-AnoGAN $\mathcal{L}_r$	0.41	0.40	0.40	0.40
c)	f-AnoGAN $\mathcal{L}_n$	0.54	0.51	0.78	0.76
d)	f-AnoGAN $\mathcal{L}_o$	0.53	0.50	0.55	0.43
e)	Ours A	0.49	0.47	0.55	0.53
f)	Ours $\mathcal{L}_r$	0.42	0.41	0.51	0.51
g)	Ours $\mathcal{L}_n$	0.58	0.56	0.78	0.78
h)	Ours $\mathcal{L}_o$	0.70	0.63	0.89	0.87
i)	Ours, proposed	**0.72**	**0.64**	**0.90**	**0.89**

f-AnoGAN fails to separate normal from anomalous samples in both $\text{CIFAR}_{\text{CAR}}$ and KTH-Cellvideos (method a and b). Method a) is the default f-AnoGAN implementation. The AUC drastically improves for KTH-Cellvideos when we add the minmax normalization to the residual loss (method c). However, the norm-based loss $\mathcal{L}_o$ cannot discriminate between normal and anomalous samples (method d).

For our method, AUC increases when we add the minmax normalization and the origin distance loss $\mathcal{L}_o$ (method g and h). The proposed method, method i), which uses a convex combination of the two achieves state-of-the art results on both KTH-Cellvideos and $\text{CIFAR}_{\text{CAR}}$.

Regarding training dataset contamination with anomalous samples, there is no degradation in AUC for the proposed method on the dataset KTH-Cellvideos, in contrast to f-AnoGAN. Some examples of closest matches for the proposed method versus f-AnoGAN can be seen in Figure 5.

5 Conclusion

In this paper, we provide an empirical study of training anomaly detectors using contaminated training data and conclude that detection performance can deteriorate. We also propose an approach to truly unsupervised anomaly detection that can maintain results even when the training data is contaminated with anomalies[3].

We conclude that *joint* generator and encoder training together with an encoder loss based on image distance is superior to training the encoder and generator separately. Joint generator and encoder training enforces similar images to lie close to each other and, thus, stratifies the latent space. At the same time, robustness to anomalies in the training data is improved.

Further work includes additional analysis of the structure of the latent space and how it is affected by different encoder losses as well as a more extensive study on the choice of the weight λ.

ACKNOWLEDGEMENTS

This research was funded by the Swedish Research Council through the project Learning Systems for Remote Thermography, grant no. D0570301, the project Energy Minimization for Computational Cameras (2014-6227), the project ELLIIT (the Strategic Area for ICT research, funded by the Swedish Government), as well as the European Union's Horizon 2020 reseach and innovation programme under grant agreement no. 783221, Aggregate FARming in the CLOUD (AFarCloud).

[3] Code is available at `https://github.com/amandaberg/GANanomalyDetection`

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

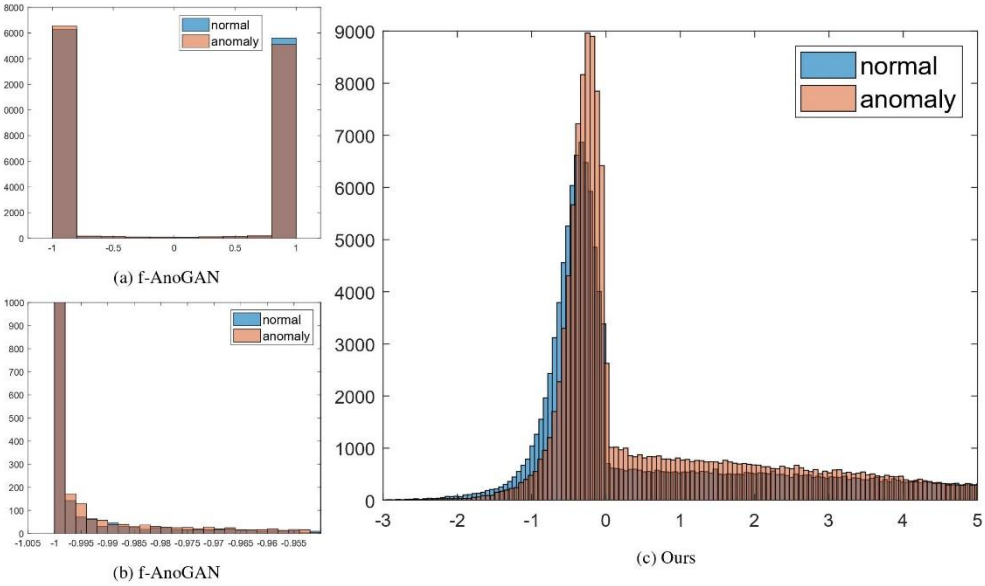


Figure 3: Histogram plots for (a) f-AnoGAN (10 bins) and (c) the proposed method (600 bins) of the coefficients of the encoded latent vectors $\hat{z}$ for the validation samples of KTH-Cellvideos. (b) shows is a plot of the same data as (a), but with different axis limits and number of bins (1000 bins).

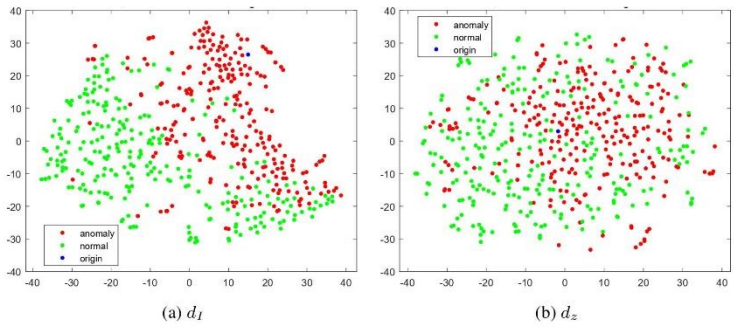


Figure 4: t-SNE visualization of validation samples projected to latent space when encoder training loss is based on the distance in (a) image space and (b) latent space.

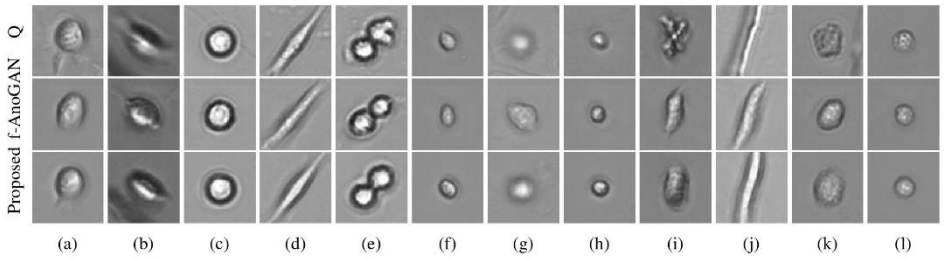


Figure 5: Closest matches for query image Q (row 1) by f-AnoGAN (row 2) and the proposed method (row 3). Columns (a)-(f) are examples of cells and columns (g)-(l) are examples of anomalies.

REFERENCES

[1] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal, 'Fraud Detection System: A Survey', *Journal of Network and Computer Applications*, **68**, 90–113, (jun 2016).

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam, 'A Survey of Anomaly Detection Techniques in Financial Domain', *Future Generation Computer Systems*, **55**, 278–288, (feb 2016).

[3] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, 'GANomaly: Semi-supervised Anomaly Detection via Adversarial Training', in *2018 Asian Conference on Computer Vision (ACCV)*, eds., C. V. Jawahar, , Hongdong Li, , Greg Mori, , and Konrad Schindler, pp. 622–637. Springer International Publishing, (dec 2019).

[4] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, 'Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection', in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (jan 2019).

[5] Jinwon An and Sungzoon Cho. Variational Autoencoder based Anomaly Detection using Reconstruction Probability, 2015.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou, 'Wasserstein GAN', *CoRR*, (abs/1701.07875), (jan 2017).

[7] Laura Beggel, Michael Pfeiffer, and Bernd Bischl, 'Robust Anomaly Detection in Images using Adversarial Autoencoders', *CoRR*, (abs/1901.06355), (jan 2019).

[8] Raghavendra Chalapathy and Sanjay Chawla, 'Deep Learning for Anomaly Detection: A Survey', *CoRR*, (abs/1901.03407), (jan 2019).

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar, 'Anomaly Detection: A Survey.', *ACM Comput. Surv.*, **41**(3), 15:1–15:58, (2009).

[10] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft, 'DeepAnomaly: Combining Background Subtraction and Deep Learning for Detecting Obstacles and Anomalies in an Agricultural Field.', *Sensors (Basel, Switzerland)*, **16**(11), (nov 2016).

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, 'Generative Adversarial Networks: An Overview', *CoRR*, (abs/1710.07035), (oct 2017).

[12] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft, 'Image Anomaly Detection with Generative Adversarial Networks', in *Machine Learning and Knowledge Discovery in Databases*, pp. 3—17. Springer International Publishing, (2018).

[13] Dit-Yan Yeung and C. Chow, 'Parzen-Window Network Intrusion Detectors', in *Object Recognition Supported by User Interaction for Service Robots*, volume 4, pp. 385–388. IEEE Comput. Soc.

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell, 'Adversarial Feature Learning', *CoRR*, (abs/1605.09782), (may 2016).

[15] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville, 'Adversarially Learned Inference', *CoRR*, (abs/1606.00704), (jun 2016).

[16] Tom Fawcett, 'An Introduction to ROC Analysis', *Pattern Recognition Letters*, **27**(8), 861–874, (jun 2006).

[17] P M Gilbert, K L Havenstrite, K E G Magnusson, A Sacco, N A Leonardi, P Kraft, N K Nguyen, S Thrun, M P Lutolf, and H M Blau, 'Substrate Elasticity Regulates Skeletal Muscle Stem Cell Self-Renewal in Culture', *Science (New York, N.Y.)*, **329**(5995), 1078–81, (aug 2010).

[18] Izhak Golan and Ran El-Yaniv, 'Deep Anomaly Detection Using Geometric Transformations', in *NIPS'18 Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9758–9769, (2018).

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative Adversarial Nets', in *Advances in Neural Information Processing Systems 27*, eds., Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, 2672–2680, Curran Associates, Inc., (2014).

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, 'Improved Training of Wasserstein GANs', in *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5769—5779, (mar 2017).

[21] Victoria J. Hodge and Jim Austin, 'A Survey of Outlier Detection Methodologies', *Artificial Intelligence Review*, **22**(2), 85–126, (oct 2004).

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, 'Progressive Growing of GANs for Improved Quality, Stability, and Variation', in *ICLR 2018*, (oct 2017).

[23] Alex Krizhevsky, 'Learning Multiple Layers of Features from Tiny Images', *University of Toronto*, (2012).

[24] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, and Kuinam J. Kim, 'A Survey of Deep Learning-Based Network Anomaly Detection', *Cluster Computing*, 1–13, (sep 2017).

[25] Klas E. G. Magnusson, Joakim Jalden, Penney M. Gilbert, and Helen M. Blau, 'Global Linking of Cell Tracks Using the Viterbi Algorithm', *IEEE Transactions on Medical Imaging*, **34**(4), 911–929, (apr 2015).

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow, 'Adversarial Autoencoders', in *International Conference on Learning Representations*, (2016).

[27] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan, 'ClusterGAN : Latent Space Clustering in Generative Adversarial Networks', *CoRR*, (abs/1809.03627), (sep 2018).

[28] Cuong Phuc Ngo, Amadeus Aristo Winarto, Connie Kou Khor Li, Sojeong Park, Farhan Akram, and Hwee Kuan Lee, 'Fence GAN: Towards Better Anomaly Detection', *CoRR*, (abs/1904.01209), (apr 2019).

[29] Emanuel Parzen, 'On Estimation of a Probability Density Function and Mode', *The Annals of Mathematical Statistics*, **33**(3), pp. 1065–1076, (1962).

[30] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks', *CoRR*, (abs/1511.06434), (nov 2015).

[31] I S Reed and X Yu, 'Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(10), 1760–1770, (1990).

[32] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth, 'f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks', *Medical Image Analysis*, 1–24, (2019).

[33] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs, 'Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery', in *Information Processing in Medical Imaging*, 146—157, (mar 2017).

[34] Waqas Sultani, Chen Chen, and Mubarak Shah, 'Real-world Anomaly Detection in Surveillance Videos', *CoRR*, (abs/1801.04264), (jan 2018).

[35] Laurens van der Maaten and Geoffrey Hinton, 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, **9**(Nov), 2579–2605, (2008).

[36] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun, 'Learning Discriminative Reconstructions for Unsupervised Outlier Removal', in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1511–1519. IEEE, (dec 2015).

[37] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar, 'Efficient GAN-Based Anomaly Detection', *CoRR*, (abs/1802.06222), (feb 2018).

[38] Houssam Zenati, Manon Romain, Chuan Sheng Foo, Bruno Lecouat, and Vijay Ramaseshan Chandrasekhar, 'Adversarially Learned Anomaly Detection', in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 727–736, (dec 2018).

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

自然场景下异常检测的无监督对抗学习

Unsupervised Adversarial Learning of Anomaly Detection in the Wild

Amanda Berg and **Michael Felsberg** and **Jörgen Ahlberg**

学　　号：	2150506021
姓　　名：	刘逸伦
导　　师：	赵季中教授
项目设计题目：	基于深度学习的工业异常检测算法设计与实现
学 科 专 业：	计算机科学与技术（少年班）
学　　院：	计算机科学与技术学院
翻 译 时 间：	2021 年 4 月 28 日

摘 要

针对高维数据（如图像等）进行异常检测任务的无监督学习是一个颇具挑战性的问题，受到了最近的大量研究。通过对正常样本的数据分布进行细致的建模，即可检测出偏离的异常样本。生成对抗网络（GAN）可以对高度复杂的正常图像样本的高维数据分布进行建模，并已被证实是解决该问题的合适方法。现有的基于 GAN 的异常检测方法通常假定存在无异常的数据可用于训练。但是，这种假设在绝大多数现实情况（即自然场景）下都是无效的。在这项工作中，本文评估了训练数据中的异常污染对于目前最先进的基于 GAN 的异常检测方法的的影响，结果检测性能如本文预期地出现了下降。为了解决这种性能下降的问题，本文提出了在训练时添加一个额外的编码器网络，阐明了生成器-编码器联合训练可以对潜空间进行分层，从而减轻了数据污染所带来的问题。本文通过实验表明，在此分层的潜空间中查询图像的范数成为区分异常与正常数据的重要线索。所提出的方法在 CIFAR-10 以及先前未经测试的细胞图像的大型数据集上均达到了最先进的性能。

1 绪论

异常检测任务是针对与正常样本相比被认为异常的稀有样本、物体或事件的识别。异常有时也称为离群值 [21]。基于任务表述的普适性，异常检测可应用于范围广泛的各种不同领域，如农业 [10]，医学 [33, 32] 和金融 [1, 2]。在机器学习的语境中，可以使用有监督学习、半监督学习或无监督学习来进行异常检测。本文正是着手于无监督的异常检测。

无监督异常检测的目的是在没有任何先验知识的情况下检测先前没有出现过的稀有物体或事件。唯一可用的信息是数据集中异常占比相当小（通常小于 1%）。由于异常在训练时对于用户来说是罕见且未知的，因此在大多数情况下，异常检测均被归结为对正常数据分布进行建模、并在该空间中定义度量以将样本分类为正常或异常的问题。在诸如图像这样的高维数据中，原始空间中对于距离的定义不再具有描述能力（维数灾难），因此需要映射到一些更合适的空间。基于隐空间，生成对抗网络（GAN）[19] 可以对复杂的高维数据分布进行建模 [11]，因此适合用来进行图像中的异常检测。与许多经典的异常检测方法相比，基于 GAN 的方法还提供了在图像中定位异常的能力 [32, 33]。尽管在最近的工作中多有进行局部的讨论 [3, 4, 12, 28, 32, 33, 37, 38]，无监督的异常检测仍是一个具有挑战性的问题。

这些已有的基于 GAN 的无监督方法的主要局限性在于，他们假设有无异常的数据可用于训练。因此，本文认为它们并不是真正的无监督学习，因为完全无异常的数据需要弱标记。GAN 训练数据的异常污染预期将会带来检测性能的降低 [7]。本文工作表明，对于目前最先进的基于 GAN 的异常检测方法 f-AnoGAN [32] 及其变体，确实存在性能降低的情况。

本文进一步使用 t-SNE 可视化 [35] 表明了验证集中的正常和异常样本分散在隐空间中，因此限制了 GAN 的分类表达能力。为了减轻这个问题，本文提出了一种与生成器联合训练的编码器，将图像映射到隐空间。联合训练加上图像距离编码器的损失，会迫使相似的图像在隐空间中也彼此靠近放置。在这个分层的隐空间中，异常样本的隐向量比正常样本具有更短的范数。本文在基于 CIFAR-10 的数据集和大型的细胞图像数据集进行的一系列实验中，经验式证明了这一点。在这两种情况下，本文的方法都可以实现最先进的性能。

1

本文的贡献：

- 本文进行了一项实证研究，以改变训练数据中异常的数量，并衡量现有异常检测方法的性能降低的程度。
- 本文提出了一种基于联合训练的编码器的、真正的无监督异常检测的方法，即使在训练数据被异常污染的情况下，该方法也可以改善结果。

2

2 相关工作

异常检测是与众多领域相关的重要任务，如恶意软件入侵检测 [24]、视网膜损伤检测 [32, 33] 以及监控视频中异常事件的检测 [34]。异常检测方法的完整综述不在本文讨论范围之内，感兴趣的读者可以参考 [8, 9]。在无监督异常检测的特定情况下，标签在训练时是未知的。本文专注于研究基于无监督深度学习的具有空间相干性的高维非序列数据（即图像）的异常检测。

关于无监督的异常检测的经典方法包括概率模型，这些概率模型可对数据分布进行建模，例如：使用非参数的核密度估计（KDE）[29]，如 [13] 所述，可以应用于入侵检测。低密度区域中的样本被视为异常。另一个概率参数化方法的示例是 RX 异常检测器（RX 算子）[31]。然而，由于维数灾难（在涉及到向量的计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象），基于概率的方法不适用于诸如图像之类的高维数据。而且，它们通常不能提供图像异常定位的功能。

与之对比的是基于重建的方法，这种方法提供了在图像中定位异常的能力。这些方法的目的是找到一个低维的隐空间，可以在该空间上重建正常样本；随后将待查询的图像投影到该隐空间上，并通过一些图像距离的测量，将重建出的图像与待查询图像进行比较，以区分异常情况。对隐空间进行建模时，可以使用例如自动编码器 [36]、变分自动编码器 [5] 或生成对抗网络（GAN）[4, 12, 32, 33, 37, 38]。在无监督的异常检测的语境下，Schlegl 等人首先引入了生成对抗网络（AnoGAN）[33]。他们建议使用 l2 范数，和衡量待查询图像与最匹配的重建图像之间的差异的判别损失函数组合，作为异常度指标。基于这种思路，Deecke 等人 [12] 提出了一种类似的方法（ADGAN），略微改善了结果：与 AnoGAN 相比，ADGAN 在隐空间中多个位置进行最接近匹配的查找。最近，Schlegl 等人 [32] 同时提出了 f-AnoGAN，通过用 Wasserstein GAN（WGAN-GP）代替深卷积 GAN（DCGAN）[30] 改进了 AnoGAN 方法，同时引入一个单独训练的编码器，来学习将图像映射至隐空间。 Zenati 等人也探索了使用编码器代替迭代的优化过程，以加速图像到隐空间的映射 [37, 38]，他们采用类似于双向 GAN（BiGAN）[14] 的架构，其中成对的 (X, z) 作为判别器的输入。与 [37，38] 相比，本文提出的方法的新颖之处在于讨论了这种编码器对隐空间结构的影响，以及训练数据污染的问题。

3

Ngo 等人 [28] 观察到，通常的 GAN 目标鼓励生成的样本分布尽可能贴合真实数据，然而这在真实的异常检测的情况下可能并非最佳策略。他们进一步提出了包围损失（encirclement loss），将生成的样本放置在分布的边界，然后可以直接使用判别器来区分异常样本。

Golan 和 El-Yaniv [18] 提出了另一种方法，该方法经过训练后可以将输入图像映射到一组几何变换。与基于重建的方法相比，该方法无法在图像中提供异常的定位。

上述的某些方法 [4, 12, 32, 33] 自称为无监督学习，但同时又假定存在可供训练的无异常数据。获取无异常数据的过程需要将正常的数据标记出来。然而，在多数真实场景中，异常对象和（或）事件相当罕见，且都难以标记。

Beggel 等 [7] 得出结论，当训练集被异常污染时，异常检测性能会降低。他们使用对抗自动编码器（Adversarial Auto Encoder）[26]，通过拒绝训练过程中潜在的异常数据来缓解问题。本文提出的方法在此种情况下以不同的方式改善了检测结果。除了拒绝以外，本文建议使用与 GAN 联合训练的编码器。正如本文在实验中展示的一样，无需在训练时拒绝异常，而应将其映射到更接近原点的位置。

4

3 方法

本文方法基于 pGAN（progressive growing GAN）[22] 和 ClusterGAN [27] 结合的架构，去除了类标签。 图 1 给出了训练和测试时的系统结构示意图。生成器和判别器与 pGAN [22] 中的相同，而编码器则受 ClusterGAN [27] 的启发。系统架构和目标函数将在接下来作进一步的描述。测试时，判别器被去除，并且生成器和编码器的参数是固定的。基于异常度指标 a，即可判定待查询的图像 Q 异常与否。

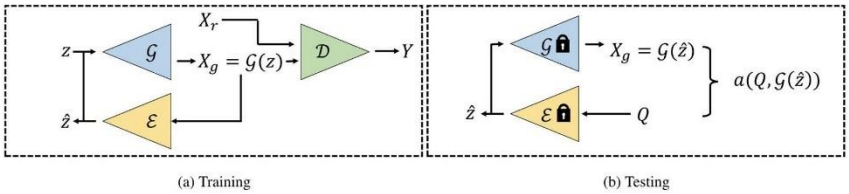


图 1 本文所提出的模型在在 (a) 训练和 (b) 测试阶段的示意图。编码器 $\mathcal{E}$ 与生成器 $\mathcal{G}$ 联合训练。在测试时，判别器 $\mathcal{D}$ 被去除，同时固定 $\mathcal{G}$ 和 $\mathcal{E}$ 的参数不变。对查询图像图像 Q 进行编码，并将其与重建 $\mathcal{G}(\mathcal{E}(Q))$ 进行比较，以求得异常度指标 a。

3.1 网络结构

AnoGAN [33] 的主要缺点之一是依赖于 DCGAN [30] 的精确重建。众所周知，DCGAN 存在模式坍塌（GAN 生成器产生的样本单一，未能拟合到完整的分布）的问题 [6]。因此，AnoGAN 的发明者在 f-AnoGAN 中使用 WGAN-GP [20] 代替了 DCGAN。不同的是，本文建议使用 pGAN（progressive growing GAN）[22]。pGAN 依然使用 WGAN-GP loss 指标，但在训练生成器和判别器时会逐渐添加新层。事实证明，这种方法可以提高 GAN 的稳定性和鲁棒性，尤其是在高分辨率图像的情况下。所提出的方法的生成器 $\mathcal{G}(z;\theta_G)$ $\mathcal{G}: z \mapsto X_g$ 和判别器 $\mathcal{D}(X;\theta_D)$ $\mathcal{D}: X \mapsto Y$ 与 pGAN 中所使用的生成器相等价。先验 $z \sim \mathcal{N}(0,1) \in \mathbb{R}^{N_z}$ 从高斯分布中得出。

与 AnoGAN 相比，f-AnoGAN 的另一创新点在于引入了编码器来代替先前的迭代搜索过程，大大提高了检测速度。编码器 $\mathcal{E}(X;\theta_E)$ 将图像映射到隐空间 $\mathcal{E}: X \mapsto \hat{z}$。与 f-AnoGAN 相比，本文所提出的方法建议将编码器 $\mathcal{E}$ 与 $\mathcal{G}$ 和 $\mathcal{D}$ 联合训练，采用与 $\mathcal{G}$、θ_G、θ_E 联合更新相同的渐进（progressive）策略。Dumoulin 等人 [15] 探索了各种编码器的训练策略，尽管是针对与本文不同的问题领域，但他们仍然强调了联合

学习 $\mathcal{G}$ 和 $\mathcal{E}$ 的重要意义。本文在实验中也进行了同样的观察。

Deecke 等人 [12] 得出了判别器不适合进行异常检测的结论。虽然在训练判别器的过程中，判别器旨在学习将真实图像与生成的图像分开、从而迫使两个概率分布重叠，但是判别器对于处理从不同分布中提取的异常样本情况则并没有得到训练。在测试时，如图 1b 所示，$\mathcal{D}$ 被去除，同时 $\mathcal{G}$ 和 $\mathcal{E}$ 以及 θ_G 和 θ_E 的参数是固定的。

3.2 目标函数

与 f-AnoGAN 和 pGAN 类似，本文采用 WGAN-GP 损失 [20]。但是，本文将 $\mathcal{E}$ 与 $\mathcal{G}$ 联合训练，而不是像 f-AnoGAN 一样在后续步骤中训练。本文所提出的方法的 GAN 目标函数采用如下形式：

$$\min_{\theta_G,\theta_E}\max_{\theta_D}\mathop{\mathbb{E}}_{X\sim p_{\text{data}}} q\big(\mathcal{D}(X)\big)+\mathop{\mathbb{E}}_{z\sim p_z} q\Big(1-\mathcal{D}\big(\mathcal{G}(z)\big)\Big)+\mathop{\mathbb{E}}_{z\sim p_z}\Big\|\mathcal{G}(z)-\mathcal{G}\Big(\mathcal{E}\big(\mathcal{G}(z)\big)\Big)\Big\|_1 \tag{1}$$

其中，由于本文使用 Wasserstein loss [27]，故 $q(x)=x$。相较于前人的工作 [20, 22, 32]，第三项 $\mathop{\mathbb{E}}_{z\sim p_z}\Big\|\mathcal{G}(z)-\mathcal{G}\Big(\mathcal{E}\big(\mathcal{G}(z)\big)\Big)\Big\|_1$ 是本文新提出的部分。

与 BiGAN 和 ALI [15] 不同的是，本文提出的架构允许 $\mathcal{G}$ 和 $\mathcal{E}$ 在训练的阶段互动，这与 ClusterGAN 所使用的编码器相类似；不同的是，ClusterGAN 在隐空间计算编码器的损失 $z-\mathcal{E}\big(\mathcal{G}(\mathrm{z})\big)$，而本文选择在图像空间 $\mathcal{G}(z)-\mathcal{G}\Big(\mathcal{E}\big(\mathcal{G}(z)\big)\Big)$ 计算编码器的损失。由 $\mathcal{G}$ 重建的待查询图像 $\mathcal{Q}$ 应当是在图像空间中距离待查询图像 $\mathcal{Q}$ 更近的匹配，而不是在隐空间中最接近的匹配，这是因为异常度指标 a（参见下一节）正是部分基于图像空间中的距离度量所定义的。此外，图像空间损失以一种不同于隐空间损失的方式构造出了隐空间，从而将正常样本和异常样本分开（参见评估部分）。

3.3 异常检测

本文建议使用由两个部分组成的异常度指标，即标准化*残差*（normalized *residual*）和*原点距离*损失（*origin distance* loss）。定义待查询图像 $\mathcal{Q}\in[0,1]^{\mathrm{W}\times\mathrm{H}\times\mathrm{D}}$ 的*残差*损失 $\mathcal{L}_{\mathrm{n}}$ 为 $\mathcal{Q}$ 与其最接近匹配 $\mathcal{G}(\hat{\mathrm{z}})$ 之间的 ℓ_2-范数：

$$\mathcal{L}_n\big(\mathcal{Q},\mathcal{G}(\hat{z})\big)=\frac{1}{N_X}\big\|w(\mathcal{Q})-w\big(\mathcal{G}(\hat{z})\big)\big\|_2 \tag{2}$$

其中 $\hat{z} = \mathcal{E}(Q)$ 是图像 Q 的编码隐向量。为了最小化图像对比度对残差损失的影响，本文与 f-AnoGAN 不同，建议采用图像的最小值－最大值标准化（minmax normalization）$w(x)$。定义标准化 $w(x): [\min(\mathrm{X}), \max(\mathrm{X})]^{W\times H\times D} \mapsto [0,1]^{\mathrm{W\times H\times D}}$（其中 $\min(\mathrm{X})$ 和 $\max(\mathrm{X})$ 查找 X 中的最小值和最大值元素）为：

$$w(x) = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{3}$$

其中除法运算是在元素层面上（element-wise）进行的，且 $N_X = W \cdot H \cdot D$。若不进行最小值－最大值标准化，较低对比度的图像样本将具有较小的残差损失，反之亦然。

基于我们对联合编码器和生成器的训练过程，以及它们如何影响隐空间结构的观察，本文将*原点距离*损失 $\mathcal{L}_o$ 定义为隐空间中从编码向量 $\hat{z}$ 到原点的距离：

$$\mathcal{L}_o(\hat{z}) = -\frac{1}{\sqrt{N_z}} \|\hat{z}\|_2 \tag{4}$$

随后定义异常度指标为 $\mathcal{L}_n$ 与 $\mathcal{L}_o$ 之间的凸组合（convex combination）：

$$a\big(Q, G(\hat{z})\big) = \lambda \mathcal{L}_n\big(Q, G(\hat{z})\big) + (1-\lambda)\big(\mathcal{L}_o(\hat{z})\big) \tag{5}$$

其中 $\lambda \in [0,1]$。如果 $a\big(Q, G(\hat{z})\big) > \alpha$，样本将被归类为异常样本。

在 [32] 中，f-AnoGAN 使用了残差损失和*判别*损失（*discrimination* loss）的凸组合作为异常度指标。判别损失取决于判别器的输出和平均判别器输出之间的差。在本文的实验中，添加这样的判别损失并未改善检测结果。

4 评估与结果

4.1 数据集

在这项工作中，本文使用了两个不同的数据集进行评估：带有完整标记的 KTH-Cellvideos 数据集 [17, 25] 为不同的细胞图像，以及 CIFAR-10 数据集 [23]。 所有训练图像均被归一化至 $[-1, 1]$ 范围内。

4.1.1 CIFAR-10 数据集

CIFAR-10 数据集 [23] 由 50000 张 $32 \times 32 \times 3$ 尺寸的 10 类训练图像（每类 5000 张图像）和 10000 张测试图像（每类 1000 张图像）组成。本文工作中使用了数据集的子集，称为 $\mathrm{CIFAR}_{\mathrm{CAR}}$：来自汽车类别的图像被视为正常样本，而来自所有其他类别的图像被视为异常样本。测试集由 1000 个正常测试样本（汽车）和 1000 个随机选择的来自其他所有类别的异常测试样本组成。

4.1.2 KTH-Cellvideos 数据集

KTH-Cellvideos 数据集 [17, 25] 由灰度的医学图像组成，这些医学图像记录了显微镜拍摄的活细胞图像序列。数据集中约有 50% 的标记对象是如气泡一类的残片（debris），并以此为类型标记。此外还标记了有丝分裂（mitosis，细胞分裂）和凋亡（apoptosis，细胞死亡）等事件，且所有细胞都有相应的遮罩。本文工作中将碎片视为异常样本，将细胞视为正常样本。

本文将数据集中的标记对象分为训练集和测试集。所有带标签的对象（正常/碎片）都被裁切至 64×64 区域范围内。此外，训练样本还被以随机角度旋转了 3 次，即原始数据集中的每个带标签的对象（为测试集保留的对象除外）产生了在训练集中的 4 个样本。总共有 $N = N_n + N_a$ 个训练样本，其中 $N_n = 525657$ 是正常训练样本的数目，$N_a = \frac{\gamma N_n}{1-\gamma}$ 是异常训练样本的数目。$\gamma \in [0, 1]$ 是由用户定义的训练数据中异常样本的百分比。测试集由 256 个正常测试图像和 256 个异常测试图像组成。

4.2 实验

为了评估所提出的方法，本文进行了一系列实验。有关网络结构和训练配置的代码以及详细说明，请访问 https://github.com/amandaberg/GANanomalyDetection。对于

8

所有实验，$N_z = 512$，$\lambda = 0.05$。在 NVIDIA GTX1080 GPU 上训练了所提出的方法，批尺寸（batch size）从 128 开始，对于 KTH-Cellvideos 数据集在 32 结束，对于 CIFAR-10 数据集在 64 结束。KTH-Cellvideos 数据集网络进行了 48 个时期（epoch）的训练（全分辨率为 6 个 epoch），CIFAR-10 数据集网络进行了 32 个 epoch 的训练（全分辨率为 4 个 epoch）。 KTH-Cellvideos 的训练时间约为 36 小时，CIFAR-10 的训练时间约为 12 小时。

所有 f-AnoGAN 网络都使用默认参数进行了训练，批尺寸为 16，z（随机向量）的尺寸为 128。KTH-Cellvideos 网络经过了 7 个 epoch 的训练。生成器的训练时间约为 16 小时，编码器的训练时间约为 1 小时。

f-AnoGAN 模型默认下的实现可以接受尺寸为 $64 \times 64 \times 1$ 的图像作为输入。$\mathrm{CIFAR_{CAR}}$ 中的图像尺寸为 $32 \times 32 \times 3$。通过将通道数增加到 3 并分别去除判别器、生成器和编码器中的一个残差块（residual block）来调整默认的模型实现。

对于数据集 $\mathrm{CIFAR_{CAR}}$，由于训练样本数量较少（5000），f-AnoGAN 生成器无法在 7 个 epoch 后生成可看的图像。即使将网络训练了多达 70 个纪元，也无法提高检测性能。因此，由于更多的迭代并没有提高检测性能，故在 $\mathrm{CIFAR_{CAR}}$ 上 f-AnoGAN 仅训练了 7 个 epoch。

异常检测结果的衡量方式为受试者工作特征曲线（Receiver Operating Characteristics (ROC)，评价分类器的分类性能最常用的指标之一）下的面积（AUC）[16]。

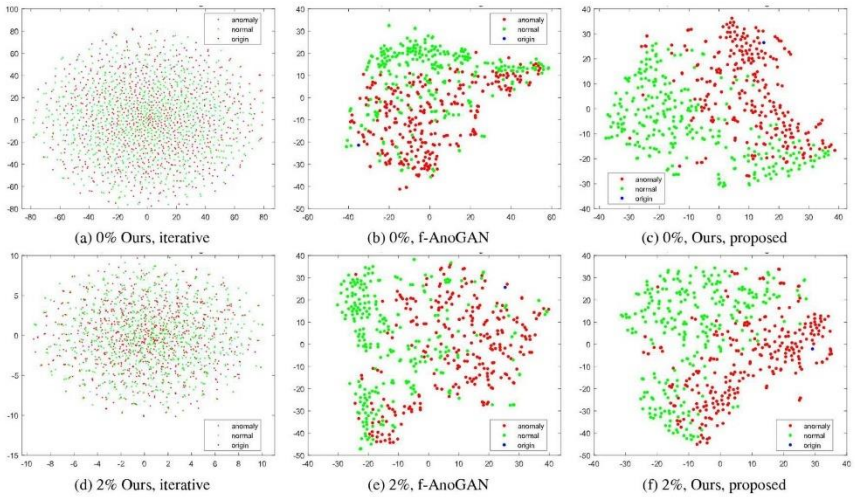
4.2.1 编码器

1）联合训练与单独训练

在 AnoGAN（注：不是 f-AnoGAN）论文 [33] 中，使用了迭代搜索来找到隐空间中与查询图像 Q 最接近的匹配项。这种方法的缺点是：a）优化可能会陷入局部最小值，并且 b）评估相当耗时。本文表明，当在没有编码器，并使用类似于 [33] 中的迭代搜索的情况下训练我们的方法时，编码后的验证样本会散布在整个隐空间中，请参见图 2a。正常和异常样本之间没有被隔开。

相反，引入编码器将隐空间进行了分层。对于 f-AnoGAN，其中编码器单独训练时，图 2b 中的样本分离（根据 t-SNE（t-Stochastic Neighbor Embedding，一种数据降维与可视化技术）似乎比图 2c 中本文所提出的方法要差一些。下文的异常度指标部

分阐述了两种方法的 AUC 指标也证实了这一点。本文认为，编码器联合训练可以使得相似的图像在隐空间中也彼此靠近。对于 t-SNE 图，使用的困惑度（perplexity value）为 30，并且在不同的运行中均采用了一致不变的可视化过程。



(a) 0% Ours, iterative (b) 0%, f-AnoGAN (c) 0%, Ours, proposed

(d) 2% Ours, iterative (e) 2%, f-AnoGAN (f) 2%, Ours, proposed

图 2 将验证集样本投影到各个方法的隐空间的 t-SNE 可视化，本文方法（a, d）在训练时没有编码器，迭代搜索最接近的匹配，（c, f）在训练时有隐空间投影的编码器以找到最接近的匹配，以及（b, e）f-AnoGAN。各个网络是在有 0%（a-c）和 2%（d-f）的异常数据比例的 KTH-Cellvideos 数据集上训练的。

图 3 中的直方图记录了 KTH-Cellvideo 数据集的验证样本的编码隐向量系数。对网络按训练数据中的异常样本率为 0%（即只使用正常样本）进行了训练。显然，本文提出的联合编码器训练可将系数更均匀地分布在隐空间上，如图 3c 所示。这些图像也解释了为什么在 f-AnoGAN 中，隐向量的范数或隐向量到原点的距离不会造成判别损失。对于 f-AnoGAN，样本最终落在超立方体中，如图 3a-b 所示。不同的是，对于本文所提出的方法（图 3c），系数在离原点近的异常处的密度更高。

接下来，我们对于为什么在经验上异常的隐变量的范数小于正常图像的范数给出了一个可能的解释。回忆 $z \sim \mathcal{N}(0,1) \in \mathbb{R}^{N_z}$。在 pGAN 的实现中，先验 z 处理之前被标准化至单位长度。从 $\mathcal{N}(0,1)$ 得出的标准化的随机向量 $z \in \mathbb{R}^{N_z}$ 将具有较小的系数。GAN 的训练会将隐空间中的数据簇从原点移开，否则判别器将无法将它们与先验的分布（即噪声）分开。编码器将正常样本映射到聚类簇中。假设异常之间的类

内变化性较高，则异常将被映射到远离聚类簇、并最终更靠近原点（即噪声），因此具有类似于随机向量的较小系数。

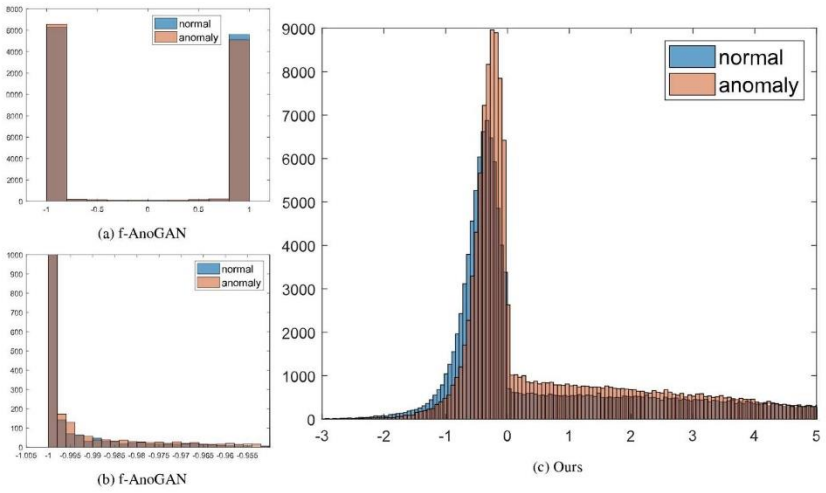


图 3　(a) f-AnoGAN（10 bins（区间））和 (c) 本文提出的方法（600 bins）对 KTH-Cellvideos 验证集样本的编码隐向量 $\hat{z}$ 的系数的直方图。(b) 显示的是与 (a) 相同的数据图，但是轴限制和区间数（1000 bins）不同。

当训练数据被异常污染时，参见图 2d 和 2e，对于 f-AnoGAN，正常样本和异常样本之间的混淆会增加。这在表 2（方法 d）中得到了证实，其中基于范数的损失 $\mathcal{L}_o$ 降低了 f-AnoGAN 的 AUC 指标。相反，即使训练数据被多达 2% 的异常（方法 h）污染，本文所提出的方法仍可保持样本之间的可分离性（图 2f）。

2）图像空间距离与隐空间距离

本文提出的编码器损失是式 (1) 中的第三项，以 d_I 表示：

$$d_I = \left\| \mathcal{G}(z) - \mathcal{G}\left(\mathcal{E}(\mathcal{G}(z))\right) \right\|_1 \tag{6}$$

将生成的图像 $\mathcal{G}(z)$ 与其在图像空间中的重建图像 $\mathcal{G}\left(\mathcal{E}(\mathcal{G}(z))\right)$ 进行比较。亦可以选择比较在隐空间中的隐向量 z 与重构的隐向量 $\hat{z} = \mathcal{E}(\mathcal{G}(z))$ 间的距离：

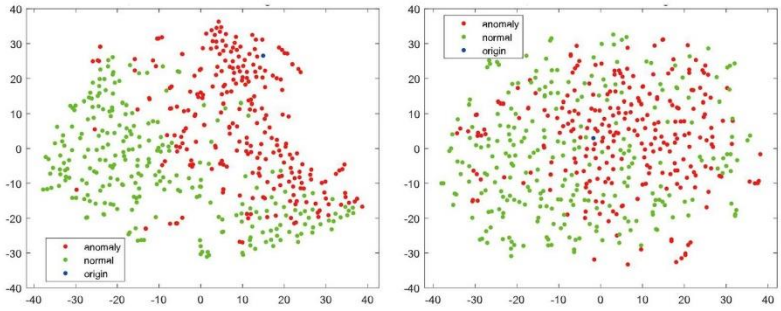
$$d_z = \left\| z - \mathcal{E}(\mathcal{G}(z)) \right\|_1 \tag{7}$$

表 1 提供了本文提出的方法在使用 d_I 和 d_z 时的结果，图 4 显示了其隐空间投

影的 t-SNE 可视化 [35]。其中网络在 KTH-Cellvideos 数据集上训练，训练数据异常率为 0%。当涉及在隐空间中区分验证集样本时，更可取的是比较图像空间距离（d_I）。好的 d_I 总是意味着好的 d_z，但反过来并非如此。我们相信这是因为 d_I 会强制（在图像空间中）相似的图像在隐空间中也彼此靠近放置，迫使重建期间 z 和 $\hat{z}$ 的小变化产生相似的图像。

表 41 对于 KTH-Cellvideos 数据集，本文提出的方法使用不同的编码器损失 d_z 和 d_I（本文提出）时的 AUC 指标结果。

编码器损失	$\mathcal{L}_n$	$\mathcal{L}_o$	$\mathcal{L}_n + \mathcal{L}_o$
d_I（本文提出）	0.78	0.89	0.90
d_z	0.66	0.69	0.66



(a) d_I (b) d_z

图 4 将验证集样本基于 (a) 图像空间距离和 (b) 隐空间距离的编码器训练损失投影到隐空间的 t-SNE 可视化。

4.2.2 异常度指标

如前所述，本文提出使用标准化残差损失 $\mathcal{L}_n$ 和基于范数的损失 $\mathcal{L}_o$ 的凸组合。在表 2 中，可以看到这些损失的不同组合在 f-AnoGAN 和本文的方法上的 AUC 指标结果。网络在两个不同的数据集的两个不同的训练数据异常比例上进行了训练。表中 A 是在 [32] 中提出的异常度指标，而 $\mathcal{L}_r$ 是同样来自 [32] 的残差损失，其中没有本文提出的最小值-最大值标准化过程。因此，

$$\mathcal{L}_r\big(Q, G(\hat{z})\big) = \|Q - G(\hat{z})\|_2 \tag{8}$$

表 2　本文所提出方法和 f-AnoGAN 使用不同异常损失和在有 0% 和 2% 异常数据比例的三个不同数据集上的 AUC 指标结果。

	方法	$\text{CIFAR}_{\text{CAR}}$		KTH-Cellvideo	
		0%	2%	0%	2%
a)	f-AnoGAN A	0.45	0.44	0.45	0.43
b)	f-AnoGAN $\mathcal{L}_r$	0.41	0.40	0.40	0.40
c)	f-AnoGAN $\mathcal{L}_n$	0.54	0.51	0.78	0.76
d)	f-AnoGAN $\mathcal{L}_o$	0.53	0.50	0.55	0.43
e)	本文方法 A	0.49	0.47	0.55	0.53
f)	本文方法 $\mathcal{L}_r$	0.42	0.41	0.51	0.51
g)	本文方法 $\mathcal{L}_n$	0.58	0.56	0.78	0.78
h)	本文方法 $\mathcal{L}_o$	0.70	0.63	0.89	0.87
i)	本文方法，提出的指标	**0.72**	**0.64**	**0.90**	**0.89**

方法 a) 和 b) 的 f-AnoGAN 无法将 $\text{CIFAR}_{\text{CAR}}$ 和 KTH-Cellvideo 视频中的正常样本与异常样本分开，其中方法 a) 是默认的 f-AnoGAN 实现。当方法 c) 将最小值-最大值标准化引入到残差损失后，在 KTH-Cellvideos 上的 AUC 指标显著提高。但是，方法 d) 基于范数的损失 $\mathcal{L}_o$ 则无法区分正常样本和异常样本。

对于本文的方法，当方法 g) 和 h) 添加最小值-最大值标准化和原点距离损失 $\mathcal{L}_o$ 时，AUC 指标会提高。所提出的方法 i) 使用了两者的凸组合，在 KTH-Cellvideos 和 $\text{CIFAR}_{\text{CAR}}$ 上均达到了目前最好的结果。

对于被异常样本污染的训练数据集，与 f-AnoGAN 相比，本文提出的方法在数据集 KTH-Cellvideos 上的 AUC 没有降低。在图 5 中可以看到 f-AnoGAN 与本文提出的方法的最接近匹配的一些示例。

13

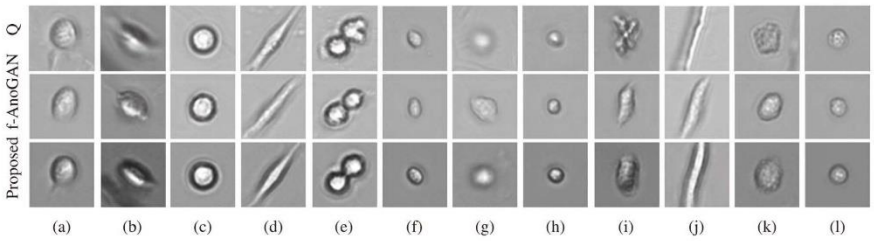


图 5：f-AnoGAN（第 2 行）和本文提出的方法（第 3 行）对查询图像 Q（第 1 行）重建得到的最接近匹配。 列 (a)-(f) 是正常细胞的示例，列 (g)-(l) 是异常的示例。

5 结论

本文实证研究了使用被异常样本污染的训练数据训练异常检测系统的方法，并得出检测性能可能会下降的结论。本文还提出了一种真正无监督的异常检测方法，即使在训练数据受到异常样本污染的情况下也可以保持良好的结果（源代码可以访问 https://github.com/amandaberg/GANanomalyDetection 获得）。

本文得出结论，*联合*训练生成器和编码器，配合基于图像距离的编码器损失，要优于分别训练编码器和生成器。生成器和编码器的联合训练会迫使相似的图像彼此靠近放置，从而对隐空间进行分层。同时，提高了对训练数据异常样本的鲁棒性。

进一步的工作包括对隐空间的结构进行额外分析、其如何受到不同的编码器损失的影响，以及对权重 λ 选择的更广泛的研究。

致 谢

这项研究是由瑞典研究委员会通过“远程热成像学习系统（Learning Systems for Remote Thermography）”项目（批号 D0570301）、“计算相机的能量最小化（Energy Minimization for Computational Cameras）”项目（2014-6227）、“ELLIIT”项目（ICT 研究战略领域，由瑞典政府资助），以及欧盟的“地平线 2020”研究与创新计划（根据编号 783221 的拨款协议）和云上的总体农业（Aggregate FARming in the CLOUD, AFarCloud）资助的。

参考文献

[1] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal, 'Fraud Detection System: A Survey', *Journal of Network and Computer Applications*, **68**, 90–113, (jun 2016).

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam, 'A Survey of Anomaly Detection Techniques in Financial Domain', *Future Generation Computer Systems*, **55**, 278–288, (feb 2016).

[3] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, 'GANomaly: Semi-supervised Anomaly Detection via Adversarial Training', in *2018 Asian Conference on Computer Vision (ACCV)*, eds., C. V. Jawahar, , Hongdong Li, , Greg Mori, , and Konrad Schindler, pp. 622– 637. Springer International Publishing, (dec 2019).

[4] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, 'Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection', in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (jan 2019).

[5] Jinwon An and Sungzoon Cho. Variational Autoencoder based Anomaly Detection using Reconstruction Probability, 2015.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou, 'Wasserstein GAN', *CoRR*, (abs/1701.07875), (jan 2017).

[7] Laura Beggel, Michael Pfeiffer, and Bernd Bischl, 'Robust Anomaly Detection in Images using Adversarial Autoencoders', *CoRR*, (abs/1901.06355), (jan 2019).

[8] Raghavendra Chalapathy and Sanjay Chawla, 'Deep Learning for Anomaly Detection: A Survey', *CoRR*, (abs/1901.03407), (jan 2019).

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar, 'Anomaly Detection: A Survey.', *ACM Comput. Surv.*, **41**(3), 15:1–15:58, (2009).

[10] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft, 'DeepAnomaly: Combining Background Subtraction and Deep Learning for Detecting Obstacles and Anomalies in an Agricultural Field.', *Sensors (Basel, Switzerland)*, **16**(11), (nov 2016).

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, 'Generative Adversarial Networks: An Overview', *CoRR*, (abs/1710.07035), (oct 2017).

[12] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft, 'Image Anomaly Detection with Generative Adversarial Networks', in *Machine*

Learning and Knowledge Discovery in Databases, pp. 3–17. Springer International Publishing, (2018).

[13] Dit-Yan Yeung and C. Chow, 'Parzen-Window Network Intrusion Detectors', in *Object Recognition Supported by User Interaction for Service Robots*, volume 4, pp. 385–388. IEEE Comput. Soc.

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell, 'Adversarial ¨ Feature Learning', *CoRR*, (abs/1605.09782), (may 2016).

[15] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville, 'Adversarially Learned Inference', *CoRR*, (abs/1606.00704), (jun 2016).

[16] Tom Fawcett, 'An Introduction to ROC Analysis', *Pattern Recognition Letters*, **27**(8), 861–874, (jun 2006).

[17] P M Gilbert, K L Havenstrite, K E G Magnusson, A Sacco, N A Leonardi, P Kraft, N K Nguyen, S Thrun, M P Lutolf, and H M Blau, 'Substrate Elasticity Regulates Skeletal Muscle Stem Cell Self-Renewal in Culture', *Science (New York, N.Y.)*, **329**(5995), 1078–81, (aug 2010).

[18] Izhak Golan and Ran El-Yaniv, 'Deep Anomaly Detection Using Geometric Transformations', in *NIPS'18 Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9758–9769, (2018).

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative Adversarial Nets', in *Advances in Neural Information Processing Systems 27*, eds., Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, 2672–2680, Curran Associates, Inc., (2014).

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, 'Improved Training of Wasserstein GANs', in *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5769–5779, (mar 2017).

[21] Victoria J. Hodge and Jim Austin, 'A Survey of Outlier Detection Methodologies', *Artificial Intelligence Review*, **22**(2), 85–126, (oct 2004).

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, 'Progressive Growing of GANs for Improved Quality, Stability, and Variation', in *ICLR 2018*, (oct 2017).

[23] Alex Krizhevsky, 'Learning Multiple Layers of Features from Tiny Images', *University of Toronto*, (2012).

[24] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, and Kuinam J. Kim, 'A Survey of Deep Learning-Based Network Anomaly Detection', *Cluster Computing*, 1–13, (sep 2017).

[25] Klas E. G. Magnusson, Joakim Jalden, Penney M. Gilbert, and Helen M. Blau, 'Global Linking of Cell Tracks Using the Viterbi Algorithm', *IEEE Transactions on Medical Imaging*, **34**(4), 911–929, (apr 2015).

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow, 'Adversarial Autoencoders', in *International Conference on Learning Representations*, (2016).

[27] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan, 'ClusterGAN : Latent Space Clustering in Generative Adversarial Networks', *CoRR*, (abs/1809.03627), (sep 2018).

[28] Cuong Phuc Ngo, Amadeus Aristo Winarto, Connie Kou Khor Li, Sojeong Park, Farhan Akram, and Hwee Kuan Lee, 'Fence GAN: Towards Better Anomaly Detection', *CoRR*, (abs/1904.01209), (apr 2019).

[29] Emanuel Parzen, 'On Estimation of a Probability Density Function and Mode', *The Annals of Mathematical Statistics*, **33**(3), pp. 1065–1076, (1962).

[30] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks', *CoRR*, (abs/1511.06434), (nov 2015).

[31] I S Reed and X Yu, 'Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(10), 1760–1770, (1990).

[32] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, ¨ and Ursula Schmidt-Erfurth, 'f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks', *Medical Image Analysis*, 1–24, (2019).

[33] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula ¨ Schmidt-Erfurth, and Georg Langs, 'Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery', in *Information Processing in Medical Imaging*, 146—-157, (mar 2017).

[34] Waqas Sultani, Chen Chen, and Mubarak Shah, 'Real-world Anomaly Detection in Surveillance Videos', *CoRR*, (abs/1801.04264), (jan 2018).

[35] Laurens van der Maaten and Geoffrey Hinton, 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, **9**(Nov), 2579–2605, (2008).

[36] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun, 'Learning Discriminative Reconstructions for Unsupervised Outlier Removal', in *2015 IEEE International*

Conference on Computer Vision (ICCV), pp. 1511–1519. IEEE, (dec 2015).

[37] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar, 'Efficient GAN-Based Anomaly Detection', *CoRR*, (abs/1802.06222), (feb 2018).

[38] Houssam Zenati, Manon Romain, Chuan Sheng Foo, Bruno Lecouat, and Vijay Ramaseshan Chandrasekhar, 'Adversarially Learned Anomaly Detection', in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 727–736, (dec 2018).

附录 B 计算机源程序