宜宾学院学报
Journal of Yibin University
ISSN 1671-5365,CN 51-1630/Z

《宜宾学院学报》网络首发论文

题目：基于 GAN 的异常检测研究综述
作者：樊富有，代洋，张淋
收稿日期：2023-01-07
网络首发日期：2023-02-16
引用格式：樊富有，代洋，张淋．基于 GAN 的异常检测研究综述[J/OL]．宜宾学院学报．https://kns.cnki.net/kcms/detail//51.1630.Z.20230215.1819.004.html

网络首发时间：2023-02-16 14:28:21
网络首发地址：https://kns.cnki.net/kcms/detail//51.1630.Z.20230215.1819.004.html

基于 GAN 的异常检测研究综述

樊富有[1,2]，代洋[3]，张淋[3]
（1.智能终端四川省重点实验室，四川宜宾 644000；2.宜宾学院 网络与图书情报信息中心，四川宜宾 644000；3.西华师范大学 电子信息工程学院，四川南充 637000）

摘　要：利用深度学习中生成式对抗网络（GAN）具有强大拟合训练数据分布能力这一最大优势，将其应用到异常检测领域可有效地准确识别异常图像。针对异常检测领域中传统有监督学习算法存在大量已知标记样本训练的局限性，以无监督学习 GAN 的异常检测模型为研究对象，阐明生成对抗网络的基本原理、网络结构及相关理论，详细介绍了近年来十种典型的基于 GAN 的异常检测模型，经过比较各衍生模型的异同，总结出各自的优势、局限性和应用场景，通过分析 GAN 在异常检测领域研究中所面临的问题及挑战，展望了未来的研究方向主要是解决模型的稳定性、计算效率、生成样本的精度、异常区域定位、异常评价机制等问题。
关键词：深度学习；生成式对抗网络；异常检测；无监督学习
中图分类号：TP391.41; TP183

Review of Research on Anomaly Detection Based on GAN

FAN Fuyou[1,2], DAI Yang[3], ZHANG Lin[3]
(1. Intelligent Terminal Key Laboratory of Sichuan Province, Yibin, Sichuan 644000, China; 2. Network and Library and Information Center, Yibin University, Yibin, Sichuan 644000, China; 3. Electronic Information Engineering, China West Normal University, Nanchong, Sichuan 637000, China)

Abstract: The biggest advantage of Generative Adversarial Network (GAN) in deep learning is its strong ability to fit the distribution of training data. Researchers apply it to the field of anomaly detection and are committed to effectively and accurately identifying abnormal images in practical applications. In the field of anomaly detection, the traditional supervised learning algorithms have the limitation of training with a large number of known labeled samples. Therefore, taking the anomaly detection model of unsupervised learning GAN as the research object to conduct a discussion. This article first introduces the basic principles, network structure and related theories of generative confrontation networks; secondly, it sorts out the typical GAN-based anomaly detection models in recent years; after that compares the similarities and differences of each derivative model, discusses and summarizes their respective superiority, limitations and application scenarios; Ultimately, the problems and challenges faced by GAN in the field of anomaly detection are discussed, and future research directions are prospected.
Keywords: deep learning; generative adversarial network; anomaly detection; unsupervised learning

近年来，深度学习（Deep Learning, DL）技术在各个领域都取得了突破性进展，并不断产出高质量的研究成果。2014 年 Goodfellow 等人[1]首次提出了生成式对抗网络（Generative Adversarial Network, GAN），因其十分擅长无监督学习任务一度引起了深度学习领域的热潮，经挖掘发现 GAN 具有巨大的研究潜力，逐渐涌现了各种各样的衍生模型，将其广泛应用于各个领域，特别是在计算机视觉领域的相关应用中成效卓著，

收稿日期：2023-01-07
基金项目：智能终端四川省重点实验室开放课题项目(SCITLAB-0019)；网络与数据安全四川省重点实验室开放课题项目(NDSZD201603)；四川省教育厅重点项目(17ZA0452)
第一作者：樊富有（1974-），男，教授，博士，研究方向为人工智能与量子计算
通信作者：代洋（1997-），男，硕士研究生，研究方向为人工智能

在技术上实现了质的飞跃。

异常检测是当下多个研究领域面临的重要问题，提高检测技术及其效率有利于开展实际应用，因而其检测速度与精度是研究的重点。异常也称为离群值，是数据中不符合明确定义的正常行为概念的模式[2]。异常检测在实际生活中应用广泛，如金融检测[3]、入侵检测[4]、欺诈检测[5]、视频监控检测[6]、生物医学检测[7]等。目前基于深度神经网络的主要研究方法有变分自动编码器（Variational Auto-Encoder, VAE）[8]、基于深度结构能量的模型（Energy Based Model, EBA）[9]和深度自编码高斯混合模型（Deep auto-encoder Gaussian mixture model, DAGMM）[10]。而针对需要大量已知标签样本的有监督学习限制了自动异常检测的能力，2017 年 Schlegl 等人[7]首次提出将 GAN 应用于异常检测领域。相较于其他生成模型，GAN 能够对真实世界数据的复杂高维分布进行建模，并且能够模拟数据分布，生成更清晰、更真实的样本，这有利于在实际异常检测中为解决类不平衡问题而进行数据增强，且在训练任务中也表现出良好的异常检测性能。至此开启了 GAN 在异常检测领域的广泛应用研究。

1 GAN 基本介绍

1.1 GAN 基本原理

GAN 主要由生成器 G（Generator）和判别器 D（Discriminator）两部分构成。判别器关注全局，生成器关注局部细节，二者相辅相成。如图 1 所示，生成器是一个用于生成类似原始图片样本的神经网络，它的输入是一个随机噪声，输出即为生成的样本。判别器是一个判别网络，它的输入是一张图片，输出的是在[0,1]范围内且代表为真实图片的概率；输出为 0 表示为生成图片，输出为 1 则表示为真实图片。

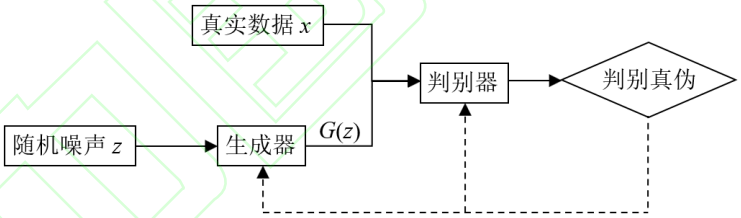

图 1　GAN 网络结构图

生成器的目的是最大程度地捕捉训练样本的特征，期望能使生成样本达到“以假乱真”的程度去欺骗判别网络。判别器的目的是将两者进行对比，尽可能分辨出输入数据的真伪，通过不断学习缩小正例样本与负例样本之间的偏差进而改善自己。在训练过程中，生成器在提高“造假能力”的同时，判别器也在提高“辨别能力”。这样，G 和 D 相互对抗就构成了一个“动态博弈”过程。在最理想的状态下，G 可以生成足以“以假乱真”的图片。对于 D 来说样本已经真假难辨，此时就达到了纳什平衡（Nash equilibrium）。

GAN 优化目标是最大化判别器参数，与此同时最小化生成器参数。其目标函数如式（1）所示：

$$\min_{G}\max_{D}V(D,G)=\underset{x\sim P_{data}(x)}{E}\left[\log D(x)\right]+\underset{x\sim P_{z}(z)}{E}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{1}$$

其中 z 表示输入的随机噪声，$P_z(z)$ 表示生成网络的分布，x 表示真实数据，$P_{data}(x)$ 表示真实数据的分布。判别器希望 $D\left(G(z)\right)$ 趋近于 0，生成器希望 $D\left(G(z)\right)$ 趋近于 1，而当 $D\left(G(z)\right)=0.5$ 时，生成样本能够以假乱真，理论上就达到了纳什平衡。

1.2 相关理论

1.2.1 WGAN 与 WGAN-GP

针对原始 GAN 在训练过程中所存在的梯度消失问题，Arjovsky 等人[11]提出了 Wasserstein 生成式对抗网络（Wasserstein Generative Adversarial Networks, WGAN）。相较于原始 GAN，WGAN 的改进之处在以下几点：在判别器最后一层去除了 sigmoid；生成器和判别器的损失函数不取 log；每次更新判别器的参数后，将它们的绝对值截断到一定范围内；推荐使用 RMSProp 和 SGD 等优化器来代替基于动量的优化算法。

WGAN 克服了 GAN 训练不稳定和模式崩塌问题，并且提高了生成样本的复杂性。另外，WGAN 使用 Wasserstein 距离取代了 JS 散度，其突出优势在于即使两个分布没有重叠部分，Wasserstein 距离仍然能够反映它们的相对分布，并且同时具有良好的平滑特性。尽管 WGAN 具有上述优点，但同时也容易产生梯度弥散或梯度爆炸现象。

针对 WGAN 存在的缺陷，Gulrajani 等人[12]提出了 WGAN-GP。WGAN-GP 用梯度惩罚（gradient penalty）代替了权值裁剪（weight clipping），在生成图像上添加了高斯噪声，用 Adam[13]优化器代替了 RMSProp。以上改进使得 WGAN 的训练变得更加稳定，并且取得更高质量的生成样本。

WGAN-GP 解决了 WGAN 的梯度爆炸和梯度消失问题，但在实际应用中发现该方法收敛缓慢，生成样本多样性欠佳。

1.2.2 BiGAN

Donahue 等人[14]针对 GAN 没有学习反向映射能力的问题提出了一种双向生成对抗网络（Bidirectional Generative Adversarial Network, BiGAN），这是一种能够同时学习样本映射与逆映射的方法。BiGAN 除基本元素外，还增加了编码器，其网络结构如图 2 所示。BiGAN 结合各个元素提出了一种优化思路：输入真实图片 x，经过编码器 E 得到 $E(x)$；从某种分布中采样随机噪声，经过解码器 G 得到 $G(z)$。通过上述两步形成了 $(x, E(x))$、$(G(z), z)$ 数据对，前者由编码器产生，后者由生成器产生。将这些数据对输入到判别器 D 中，让其辨别数据来源，这是一个双向学习的过程。

BiGAN 的出现使得 GAN 能够具备表征学习的能力，其训练稳定，可生成高质量图像，为之后的 GAN 异常检测的部分衍生模型奠定了基础。

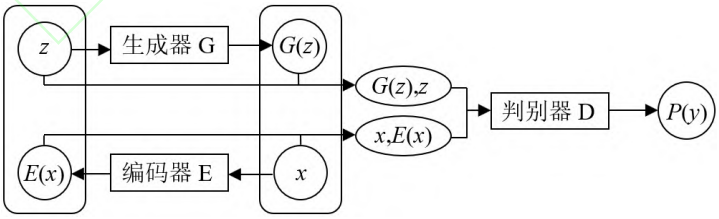

图 2　BiGAN 网络结构图

2　典型的 GAN 异常检测模型

2.1 AnoGAN 与 f-AnoGAN

2.1.1 AnoGAN

Schlegl 等人[7]针对在疾病诊断过程中，需要耗费大量的人力和时间对数据进行标注的有监督异常检测，限制了利用成像数据进行治疗决策能力的问题，提出了一种基于无

监督学习的深度对抗网络模型 AnoGAN（Anomaly Detection GAN），首次将 GAN 应用于异常检测领域。

AnoGAN 模型的思想是 GAN 只学习正例的分布，因此重构图像与输入图像特征分布应大致相同，然后通过二者之间的残差确定异常图像。AnoGAN 的异常检测框架包括两个部分：GAN 模型训练部分和异常检测部分（如图 3）。训练阶段 Schlegl 等人[7]首先对提取的正常 2D 图像块进行预处理，将提取的视网膜区域图像进行灰度值归一化至[-1,1]范围内，并将其扁平化以调整方向、形状和厚度，然后通过自动分割算法找到视网膜的顶层和底层，共提取 1000000 个 64×64 像素的 2D 图像作为 GAN 模型训练的输入。测试阶段采用包含正常样本和病理样本的 8192 个图像块作为测试集。GAN 模型采用的是深度卷积生成对抗网络[15]（Deep Convolutional Generative Adversarial Networks，DCGAN），生成器和判别器都由 4 层卷积神经网络构成，最终通过动态博弈达到平衡。

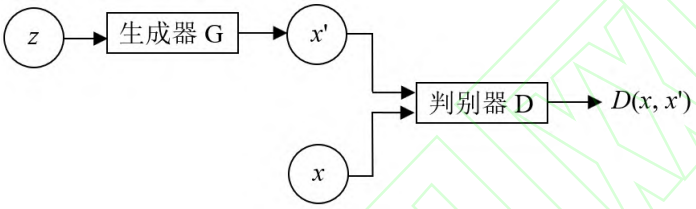

图 3　AnoGAN 网络结构图

基于视网膜临床高分辨率 SD-OCT 扫描中提取的标记测试数据集，从三个层面说明了 AnoGAN 模型异常检测的有效性：能否生成逼真图像、异常检测的准确性和异常评分性能。尽管 AnoGAN 表现出良好的异常检测性能，但在实际应用过程中该模型需要不断迭代优化，每次一个新的图像的异常检测都需要不断反向传播更新 z，势必会耗费大量时间，表现出计算效率低的缺点。

2.1.2　f-AnoGAN

针对有监督学习需要耗费大量数据且只能处理训练样例中存在的情况，同时为了提高训练速度，Schlegl 等人[16]在 AnoGAN 模型的基础上进一步提出了 f-AnoGAN（fast-AnoGAN）模型，将从图像到潜在空间的学习映射取代了 AnoGAN 中的迭代过程，极大地提高了计算速度。他们提出了一种编码器，可以快速地将图片映射到隐空间中的某个点，然后利用 WGAN[11]进行异常检测。其异常检测框架分为两个部分：模型训练和异常检测；其中模型训练部分包括 WGAN 和编码器的训练。在视网膜光学相干断层扫描正常图像数据集上训练了模型，训练完毕后参数不再改变，然后训练由卷积神经网络（Convolutional Neural Network, CNN）构成的编码器，它负责将训练图片映射为隐空间的向量 z，生成器将向量 z 映射为图片 $G(z)$，正常情况下生成图像与输入图像应相近，异常情况下二者差距较大且大于某个值。其使用的数据集及预处理方法与文献[7]一致。Schlegl 等人提出三种不同的训练方式：izi、ziz 和 izi_f，经性能比较选择 izi_f，即通过比较判别器中间某一层的特征图得出差异，网络结构如图 4 所示。

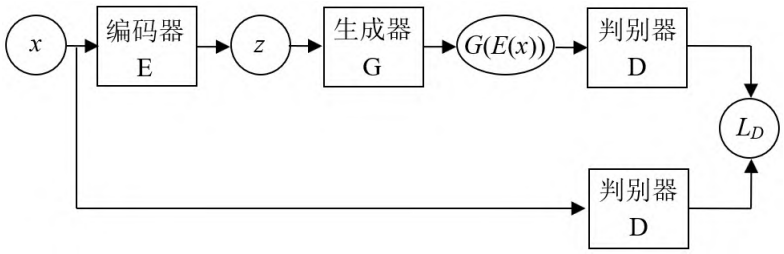

图 4　izi$_f$ 网络结构图

基于视网膜光学相干断层扫描数据集，将 f-AnoGAN 与 AE、AdvAE、ALI 模型、AD（使用 WGAN 的判别器输出作为异常分数）和 iterative（使用 WGAN 的 AnoGAN 模型）进行对比，文献[16]分别进行了隐空间平滑度、模型预测准确率、异常检测效果和不同编码器训练策略四类实验的对比研究，结果表明，编码器的 izi$_f$ 策略的训练结果最佳。此外，f-AnoGAN 模型以平滑的表示方式捕获正常样本的可变性，模型表现相当优异，能够准确地检测出异常图像，并且可以定位图像中的异常。

因此，文献[16]指出 f-AnoGAN 模型一般适用于各种生物医学数据的异常检测，其主要应用领域在于区分正常和异常图像，并以粗略异常定位的形式提示相应的异常区域，对异常区域识别具有一定的提示和参考作用。但该方法仍然存在一定的局限性：对异常检测分割精度的定量评估只是一个粗略的指示，表明它可以定位异常，不能精确到局部；从生物标记物发现的角度来看注释仅涵盖部分可能的异常。

2.2　EGBAD

Zenati 等人[17]针对 AnoGAN 模型的耗时问题做了相关改进，提出一种基于 BiGAN[14]的异常检测模型——EGBAD，其模型如图 5 所示。在训练时，同时学习编码器、生成器和判别器，如此可避免测试时迭代反向传播的耗时问题。此外，与 AnoGAN 的不同之处在于，将真实数据经编码器产生的 z' 和 z 经生成器产生的 x' 均输入到判别器中，判别器 D 不仅考虑输入，还考虑了潜在表示。

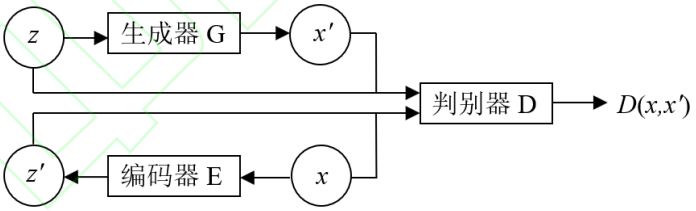

图 5　EGBAD 网络结构图

基于 MNIST 数据集[18]和 KDD99 数据集[19]的实验证明了 EGBAD 模型在高维复杂数据集上良好且高效的性能，此方法在测试时间上相较于 AnoGAN 可快数百倍。

2.3　ALAD

Zenati 等人[20]提出了一种基于 BiGAN 的逆向学习异常检测方法 ALAD 即对抗学习异常检测，该方法为异常检测任务导出对抗学习特征，然后 ALAD 利用基于这些对抗学习特征的重构误差来确定数据样本是否异常，其网络结构如图 6 所示，除基本元素外，还结合使用了三个判别器 D_{zz}、D_{xz} 和 D_{xx}。判别器 D_{xz} 用于保证模型的循环一致性，两个对称的判别器 D_{zz} 和 D_{xx} 作为条件熵约束来稳定 GAN 训练，此外还引入了光谱归一化。ALAD 使用一种基于重构的异常检测技术用于评估输入样本与重构样本之间的距离，即

计算循环一致性判别器 D_{xx} 的特征空间中样本之间的距离。ALAD 作为 EGBAD 的加强版确保了循环一致性和训练稳定性，从而显著提高了异常检测性能。

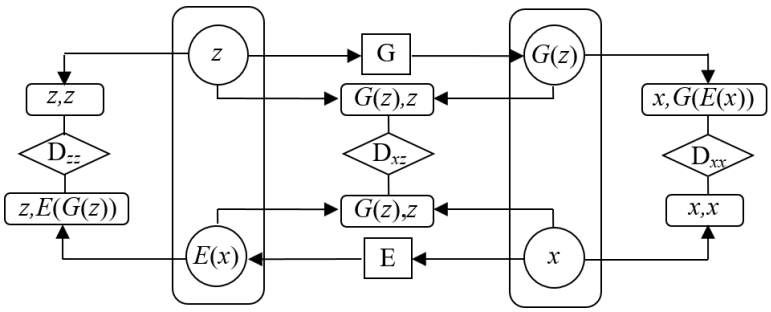

图 6　ALAD 网络结构图

基于 KDDCup99 10%、SVHN[21]、CIFAR-10[22]和 Arrhythmia 数据集，通过与 OC-SVM[23]、IF[24]、DSEBM[9]、DAGMM[10]和 AnoGAN 对比实验和消融实验表明，ALAD 优于其他检测方法，且比 AnoGAN 快几个数量级。综上可见 ALAD 模型对于复杂高维数据异常检测的有效性，表明该方法具有较强竞争力。

2.4　GANomaly 与 skip-GANomaly

2.4.1　GANomaly

Akcay 等人[25]为检测未知的异常情况，引入了一种新的网络布局，在生成器网络中使用编码器-解码器-编码器，使其能够将输入图像映射到低维向量，然后使用该低维向量重构得到输出图像，最后使用附加编码器网络将生成的图像映射到其潜在表示。

GANomaly 的检测框架如图 7 所示，主要由生成器、判别器和编码器构成，其中生成器采用蝴蝶结式的自动编码器网络构成，输入样本 x 经编码器压缩后得到潜在空间的特征向量 z，而后由 DCGAN 构成的解码器网络重建图像 x'，最后通过编码器得到重构后的特征向量 z'。D 是文献[15]中 DCGAN 引入的标准判别器网络，在训练期间最小化图像和潜在向量之间的距离。

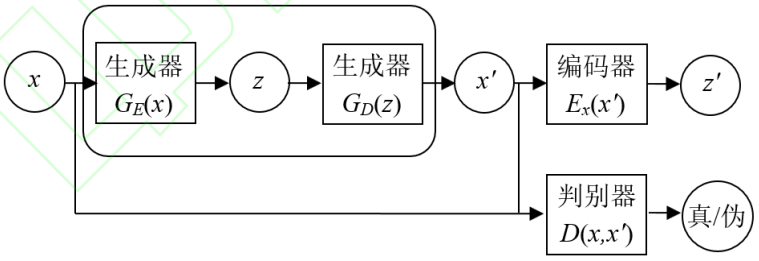

图 7　GANomaly 网络结构图

仅正常样本训练好的生成器只能拟合正常数据分布，因未训练负例样本，模型对负例的分布是未知的，故将其用于未知的异常样本检测理论上会得到差距较大的潜在空间向量 z。当两次编码的差值大于预设阈值时，即可判定该样本为异常样本。

基于 MNIST、CIFAR-10 和 X 射线安全筛查（UBA 和 FFOB）[26]数据集，并在 X 射线安全筛查的异常检测环境中进行试验，通过 GANomaly 模型与 AnoGAN、VAE 和 EGBAD 模型进行对比，发现 GANomaly 模型对异常样本具有较强的鉴别能力，该模型比其他三种方法更有效、更优越，模型的优点在于具有对任何异常检测任务的泛化能力。GANomaly 在无负例训练的情况下实现了异常检测，这在实际生活应用中具有重要意义。

虽然潜在向量包含了充分的正例样本特征，但并不能确定这些特征是否是必要特征，因此可能会导致某些异常能被重建；此外该模型也存在训练过程不稳定，每轮训练的精度变化较大，容易出现梯度弥散等问题。

2.4.2 Skip-GANomaly

Akcay 等人[27]针对 GANomaly 模型的缺陷以及数据的类不平衡问题，进一步提出了一个无监督异常检测模型——Skip-GANomaly，该模型采用带跳跃连接的 U-net[28]卷积神经网络，使网络具有更强的重建能力，其网络结构如图 8 所示。

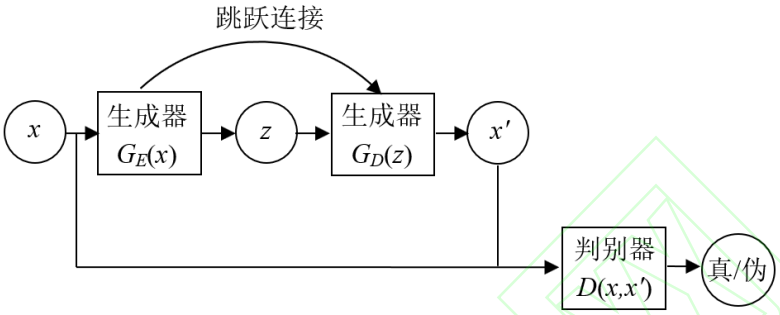

图 8 Skip-GANomaly 网络结构图

生成器由“蝴蝶结”式的 U-net 卷积神经网络构成，包括编码器和解码器。网络通过包含卷积层和 BatchNorm 层以及 LeakyReLU 激活函数的五个块读取输入 x，向下采样输出潜在表示 z，这也被称为瓶颈特征。然后由对称网络将潜在向量向上采样到输入图像维度，并重建输出，表示为 x'。生成器的任务是重建数据。判别器的任务是预测输入数据的类别，不仅能使生成器学习正常的数据分布，还能作为特征提取器学习低维潜在空间内的正态分布。模型仅对正例样本进行训练，以捕获高维图像空间中正常样本的多尺度分布。文献[27]提出的异常评分不仅包括输入图像与重建图像的差值，还包括其对应潜在特征的差值。引入潜在特征损失是为了确保网络能够为常见示例生成上下文合理的潜在表示。

实验基于 CIFAR-10、UBA 和 FFOB 数据集，研究结果表明，带跳跃连接的 U-net 网络提供了更稳定的训练，并且与 AnoGAN、EGBAD 和 GANomaly 相比取得了更好的结果。Skip-GANomaly 模型的突出特征在于编码器网络中的每个下采样层都与相应的上采样解码器层相连接，这种跳跃连接的使用为对应卷积层之间的直接信息传输提供了实质性的优势，同时也保留了图像的多尺度信息，从而得到了更好的重构。此模型对任何异常检测任务具有同样的泛化能力。但该模型可能由于存在忽略图像底层特征的情况而导致异常漏检，也有一定局限性。

2.5 保证一致性的 BiGAN

Komoto 等人[29]针对 BiGAN 存在图像空间和潜在空间之间的相互映射缺乏一致性、难以通过图像的潜在变量进行调节的问题，提出了一种保证一致性的双向 GAN 异常检测模型，该模型能够同时学习潜在空间与图像空间的双向映射。

模型在 BiGAN 基础上增加了额外的编码器和生成器，同时引入了一致性损失以确保相互映射的一致性。此外用投影判别器取代普通判别器，从而有效地集成图像和潜在变量，生成更精确的图像。其网络结构如图 9 所示。

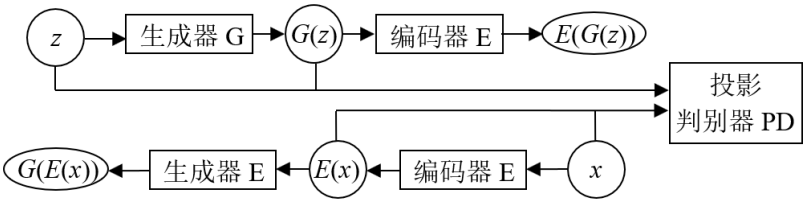

图 9　保证一致性的 BiGAN 网络结构图

实验基于 MNIST 数据集和 MVTec 异常检测数据集[30]，为证明模型的有效性，与 EGBAD、AutoEncoder[31]和 AnoGAN 模型进行了对比实验。实验结果表明，一致性损失的引入解决了图像空间与潜在空间的相互映射问题，投影判别器的引入提高了重建图像的质量和异常检测的准确性。综上所述，该模型能够检测异常，且优于传统基于 GAN 的检测方法。

2.6　DED-GAN

Liu 等人[32]为解决工业环境中机器人的异常检测问题，提出了一种具有双编码器-解码器器生成对抗网络（DED-GAN）的异常检测模型，训练时不需要任何负例样本就可检测异常。采用 DED-GAN 方法将高维输入图像映射到低维空间，通过该空间得到潜在变量，最后通过计算两个编码器得到的两个低维向量之间的距离来实现异常检测。

DED-GAN 模型网络架构如图 10 所示。生成器网络中包含两个类似的子结构，每个子结构由编码器-解码器组成。编码器对输入图像进行降维，其网络由批量归一化层[33]（Batch normalization）和 ReLU 激活函数构成。解码器对图像进行重构，其网络由反卷积层、批归一化层和 ReLU 激活函数构成。判别器由 DCGAN 构成。在训练阶段，生成器和判别器通过不断竞争优化网络，直至收敛。

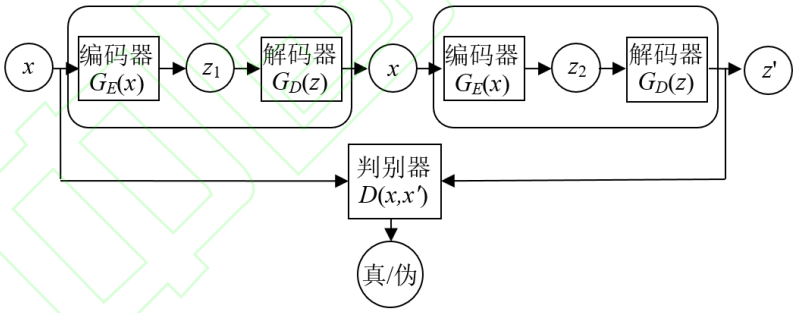

图 10　DED-GAN 网络结构图

文献[32]的创新点在于，设计了一种基于 DED 式的新架构，用于捕获训练过程中的图像分布，以便准确识别异常样本。此外，引入了一种新的权值函数来控制编码重建和对抗阶段的损失比例，以生成高质量图像。实验基于 CIFAR-10、MNIST 和 Kolektor 表面缺陷数据集[34]，结果表明与 VAE、EGBAD、AnoGAN 和 GANomaly 模型相比，DED-GAN 具有更优的性能。该方法适用于异常样本很少或难以获得的应用场景。

2.7　MAD-GAN

Dan 等人[35]研究了有关时间序列方面的异常检测，传统的基于阈值方法和有监督机器学习方法暴露出系统的动态复杂性不足和缺乏大量标记数据的缺点，而无监督学习方法未充分利用多变量的时空相关性。因而文献[35]探讨了如何使用 GAN 对网络物理系统（Cyber-physical systems, CPSs）中的多变量时间序列数据进行多元异常检测，提出一

种 MAD-GAN 的多元异常检测框架，该框架同时考虑整个变量数据集，以捕获各变量之间的潜在交互。

MAD-GAN 的网络框架如图 11 所示，主要由 GAN 模型训练和序列异常检测两部分组成。每部分的生成器 G 和判别器 D 都由长短期记忆循环神经网络（Long Short Term-Recurrent Neural Networks, LSTM-RNN）构成。模型利用滑动窗口将多变量时间序列分成多个等长的子序列，每个序列即为一个样本，然后通过迭代对抗训练生成器和判别器，达到平衡后参数保持不变以便异常检测时使用。异常检测利用 G 和 D 使用判别和重建异常评分（DR-Score）来检测异常。

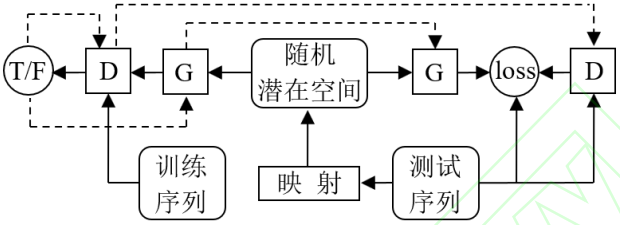

图 11　MAD-GAN 网络结构图

基于 SWaT 和 WADI 两个复杂网络攻击 CPS 数据集进行的测试表明，与 PCA[36]、KNN[37]、FB[38]和 AE[39]四种检测方法进行对比，在复杂的现实网络物理系统中，模型对于由各种网络入侵引起的异常有效，并显示出优于以上无监督检测方法的性能。文献[35]的主要创新点在于提出了基于 GAN 的多变量时间序列异常检测模型 MAD-GAN；且 D 和 G 使用 LSTM-RNN 来捕获时间上的依赖关系；分别利用 D 和 G 计算判别和重建损失，进而计算 DR 异常评分，然后进行异常检测；最后开源了两个数据集：安全水处理系统（Secure Water Treatment, SWaT）和水分配（Water Distribution, WADI）数据集。

2.8　GAN 集成

Han 等人[40]在实验中观察到 GAN 集成在生成任务中通常优于单个 GAN，于是将多个生成器（由编码器-解码器构成）与多个判别器相结合，提出了基于 GAN 集成的异常检测方法。在对抗训练中，判别器对来自各个生成器的样本与原始样本进行判别更新，但每次只迭代更新一个生成器-判别器对。

文献[40]在运行时间和性能二者的权衡之下，选择三对生成器和判别器进行实验。GAN 集成网络结构如图 12 所示，原始图像被输入到多对编码器-解码器中，经过编码器的压缩和解码器的重建后，得到重构图像，然后多对重构图像和原始图像都被输入到对应判别器中，最后通过编码器-解码器-判别器计算出异常评分的平均值，平均值的改进有利于消除虚假分数。

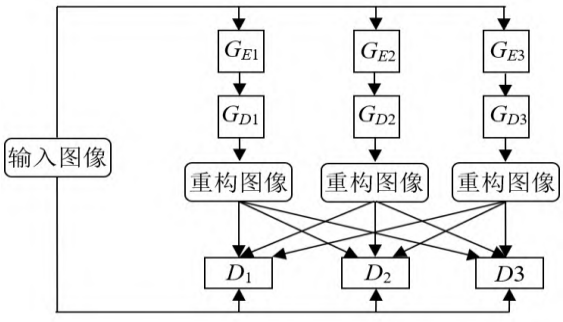

图 12　GAN 集成网络结构图

基于KDD99、OCT[41]、MNIST和CIFAR-10数据集，考虑了四个基本模型：f-AnoGAN、EGBAD、GANomaly和Skip-GANomaly，将这四个基本模型与对应的集成模型相比较，实验表明，GAN 集成模型在异常检测任务中明显优于单一模型，集成有助于捕获数据模式，并且能够更细致的模拟数据分布，同时能够显著提高模型效能。

文献[40]从潜在向量和编码向量两者间分析了 GAN 集成能显著提高性能的原因：多个生成器有助于捕获数据模式，并提供具有多样性的合成样本和更大的联合支持，另外，集成也主要改善了 GAN 的判别损失。由此可见，多个生成器比单一模型能训练出更好的判别器。

3 GAN 异常检测模型比较研究

表 1 总结了近年来以 GAN 为基础的十种较为典型的异常检测模型，分别介绍了每个模型的运作机制，并对比了各自的优势、局限性及应用场景。其中具有代表性和开创性的是 AnoGAN，它首次将 GAN 进行改进并运用于异常检测领域中，提出将数据映射至潜在空间的映射方式，并定义置信分数以衡量异常。另一个具有较大影响力的是 EGBAD，主要针对 AnoGAN 计算效率低的缺陷进行改进，提出基于 BiGAN 的双向映射，并能同时训练编码器、生成器和判别器的模型，极大地提高了运行速度。

表 1　典型 GAN 异常检测模型的总结与对比

GAN 衍生模型	运作机制/创新点	优势	局限性	适用场景
AnoGAN	通过重构图像与输入图像之间的残差来确定异常	首次将 GAN 应用于该领域，具有良好的检测性能	代价昂贵，计算效率低下，没有考虑到逆映射	生物医学疾病诊断
f-AnoGAN	使用判别器引导的从图像到潜在空间的学习映射	能够平滑的捕获正常可变性，具有较好的异常检测和分割性能	只能对图像异常进行粗定位，不能精确到局部	应用于实时异常检测，如各种生物医学异常检测
EGBAD	同时学习编码器、生成器和判别器，且判别器同时考虑了输入图像及潜在表示	相较于 AnoGAN，运行时间可快数百倍	生成样本质量不高，特征表示的学习效果不佳，异常评分的可解释性差	大部分图像的异常检测
ALAD	使用了三个判别器以保证模型的循环一致性和训练稳定性	比 AnoGAN 快几个数量级，加强了模型的稳定性，此外提出新的置信度	在图像数据的某些异常检测任务中表现不佳	适用于表格数据上的异常检测任务
GANomaly	生成器使用编码器-解码器-编码器的对抗网络结构，还引入了上下文损失	无需在负例上训练，新的架构使整个学习过程更快	存在图像空间与潜在空间检测结果无法匹配的情况	运输过程中的 X 射线安全检查
Skip-GANomaly	生成器使用带跳跃连接的 U-net 卷积神经网络	可捕获高维图像的多尺度分布，具有更强的重建能力	可能存在忽略图像底层细节特征而导致漏检	运输过程中的 X 射线安全检查
保证一致性的 BiGAN	引入了一致性损失，使用投影判别器取代普通判别器	保证了映射及逆映射的一致性，能够有效整合图像及其	部分缺陷偏小影响检测准确率，实际工业应用可能存在一定差距	适用于工业环境中电子器件表面瑕疵检测，如金属螺母异常检测

		潜在向量，并生成更精确的图像		
DED-GAN	使用双编码器-解码器对抗网络，引入了新的权值函数来控制损失比例	可准确识别异常样本并生成高质量图像	对部分图像或数字识别存在一定局限	适用于异常样本很少或难以获得的情况，如实际工业环境中机器人的异常检测
MAD-GAN	使用由 LSTM-RNN 构成的生成器和判别器，提出 DR-Score 异常评分	擅长捕获时间上的依赖关系，提出新的异常评分，并开源了两个数据集	无法确定最佳子序列长度，模型训练不稳定	适用于网络物理系统中的多变量时间序列数据检测
GAN 集成	多对生成器与判别器结合进行对抗训练，异常评分取其平均值	在异常检测任务中明显优于单一模型	异常分数取平均值使得检测性能有限	大部分图像异常检测

部分模型之间也存在共性，AnoGAN、f-AnoGAN、EGBAD、GANomaly、Skip-GANomaly 和 DED-GAN 这些模型结构类似，都使用编码器-解码器作为生成器，编码器用于压缩图像即进行降维，解码器用于逆映射以重构图像，特别是 DED-GAN 使用了双编码器-解码器，能够更准确的捕获图像分布。保证一致性的 BiGAN 和 EGBAD 模型都是在 BiGAN 基础上延伸扩展而成。此外，MAD-GAN 则使用由 LSTM-RNN 构成的生成器和判别器，以便处理时间序列数据来捕获时间依赖性。ALAD 除生成器和编码器外引入了三个判别器来保证模型的循环一致性和训练稳定性。

近年来，针对不同异常检测的应用需求，衍生出不同的 GAN 异常检测模型，如结合注意力机制[42]、一致性约束[43]等着重于改进模型的训练速度、生成样本精度和判别器置信度。这些模型在实践中各有优劣，在实际应用中应针对不同的检测对象和检测需求选择合适的检测模型。

4 研究展望

自异常检测领域首次应用 GAN 以来，短短几年时间就取得了突破性进展，产生众多的 GAN 检测模型，其未来发展可谓是日新月异。就本文分析讨论的 GAN 异常检测衍生模型来看，它们在异常检测领域都有极大的研究潜力和应用价值，但同时也面临着多重挑战，这也是该领域中生成对抗网络的未来研究方向。

（1）模型稳定性问题。自生成对抗网络开创以来，就存在训练稳定性差的问题，GAN 的异常检测模型也不例外。尽管研究者们进行了部分改进，但这仍是目前需要克服的问题。要想达到理论上的纳什平衡，还需进一步研究。

（2）模型计算效率问题。目前一些 GAN 异常检测模型为达到良好的检测性能，以运行时间为代价，导致模型计算效率低下。面对大量的异常检测样本，极长的计算时间是不现实的，未从实际出发考虑二者如何兼顾问题。因此，计算效率有待提高。

（3）生成样本的精度问题。在异常检测中大多采用从图像空间到潜在空间的相互映射来重构样本，然而模型在训练过程中存在损失，这可能导致重构样本与输入图像之间存在细节丢失，以至于生成样本的质量不高。因此，生成样本的精度问题有待进一步改善。

（4）异常区域定位问题。目前大部分异常检测方法均只能检测异常样本，虽然文献[16]所提出的方法能够定位异常，但只能粗略定位，且这方面少有论文做相关研究。若能精确定位异常，则能及时得知异常点，并做出相关改进，这在实际应用中具有重大研究意义。因此，对于异常定位问题，有待提出更为优质的检测模型。

（5）异常评价机制问题。从以上典型 GAN 异常检测模型研究来看，大多采用重构特征向量与输入图像潜在向量之间的差值作为异常衡量标准，但这种评价方式过于单一，应考虑置信度的可解释问题。因此，完善丰富的评价机制有待进一步研究。

参考文献：

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

[2] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[3] Ahmed M, Mahmood A N, Islam R. A survey of anomaly detection techniques in financial domain[J]. Future Generation Computer Systems, 2016(55): 278-288.

[4] Atefeh F, Khreich W. A Survey of Techniques for Event Detection in Twitter[J]. Computational Intelligence, 2015, 31(1): 132-164.

[5] Abdallah A, Maarof M A, Zainal A. Fraud detection system: A survey[J]. Journal of Network and Computer Applications, 2016(68): 90-113.

[6] Kiran B R, Thomas D M, Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos[J]. Journal of Imaging, 2018, 4(2): 36.

[7] Schlegl T, Seeböck P, Waldstein S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]//International Conference on Information Processing in Medical Imaging. Springer, Cham, 2017: 146-157.

[8] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability[J]. Special Lecture on IE, 2015, 2(1): 1-18.

[9] Zhai S, Cheng Y, Lu W, et al. Deep structured energy based models for anomaly detection[C]//International conference on machine learning. ACM, 2016: 1100-1109.

[10] Zong B, Song Q, Min M R, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection[C]//International Conference on Learning Representations. 2018-01-30. [2022-07-17]. https://openreview.net/forum?id=BJJLHbb0-.

[11] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International Conference on Machine Learning. ACM, 2017: 214-223.

[12] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs[C]//International Conference on Neural Information Processing Systems. ACM, 2017: 5769–5779.

[13] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[EB/OL]. 2014-12-22. arXiv:1412.6980.

[14] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning[EB/OL]. 2016-05-31. arXiv:1605.09782.

[15] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. 2015-11-19. arXiv:1511.06434.

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

[17] Zenati H, Foo C S, Lecouat B, et al. Efficient GAN-Based Anomaly Detection[EB/OL]. 2018-02-17. arXiv:1802.06222.

[18] LeCun Y, Cortes C. The mnist database of handwritten digits[EB/OL]. [2022-07-17]. https://www.semanticscholar.org/paper/The-mnist-database-of-handwritten-digits-LeCun-Cortes/dc52d1ede1b90bf9d296bc5b34c9310b7eaa99a2.

[19] Lichman M. UCI Machine Learning Repository, 2013[DB/OL]. http://archive.ics.uci.edu/ml/index.php.

[20] Zenati H, Romain M, Foo C S, et al. Adversarially learned anomaly detection[C]//2018 IEEE International Conference on Data Mining (ICDM). Singapore: IEEE, 2018: 727-736.

[21] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[J]. Nips, 2011: 1-9.

[22] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4).

[23] Schölkopf B, Williamson R C, Smola A J, et al. Support vector method for novelty detection[C]// International Conference on Neural Information Processing Systems. ACM, 1999: 582-288.

[24] Liu F T, Ting K M, Zhou Z H. Isolation forest[EB/OL]. [2022-07-21]. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf.

[25] Akcay S, Atapour-Abarghouei A, Breckon T P. GANomaly: Semi-supervised anomaly detection via adversarial training[C]//Asian conference on computer vision. Springer, Cham, 2019: 622-637.
[26] Akcay S, Kundegorski M E, Willcocks C G, et al. Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(9):2203-2215.
[27] Akçay S, Atapour-Abarghouei A, Breckon T P. Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection[C]//2019 International Joint Conference on Neural Networks (IJCNN). IEEE, Budapest: IEEE, 2019: 1-8.
[28] Ronneberger O, Fischer P, Brox T. U-NET: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2015: 234-241.
[29] Komoto K, Aizawa H, Kato K. Consistency ensured bi-directional GAN for anomaly detection[C]//International Workshop on Frontiers of Computer Vision. Springer, Singapore, 2020: 236-247.
[30] Bergmann P, Fauser M, Sattlegger D, et al. MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9592-9600.
[31] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[32] Liu H, Tang T, Luo J, et al. An anomaly detection method based on double encoder–decoder generative adversarial networks[J]. Industrial Robot, 2021, 48(5): 643-648.
[33] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. ACM, 2015: 448-456.
[34] Tabernik D, Šela S, Skvarč J, et al. Segmentation-based deep-learning approach for surface-defect detection[J]. Journal of Intelligent Manufacturing, 2020(31): 759–776.
[35] Li D, Chen D, Jin B, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks[C]//International Conference on Artificial Neural Networks. Springer, Cham, 2019: 703-716.
[36] Li S, Wen J. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform[J]. Energy and Buildings, 2014, 68(A):63-71.
[37] Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces[C]//European Conference on Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer, 2002: 15-27.
[38] Lazarevic A, Kumar V. Feature bagging for outlier detection[C]//Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. ACM, 2005: 157-166.
[39] Han J, Kamber M, Pei J. Data mining: Concepts and techniques[M]. San Francisco: Morgan Kaufinann, 2006: 559-569.
[40] Han X, Chen X, Liu L P. GAN ensemble for anomaly detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(5): 4090-4097.
[41] Kermany D S, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning[J]. Cell, 2018, 172(5): 1122-1131.e9.
[42] Li X, Zheng Y, Chen B, et al. Dual attention-based industrial surface defect detection with consistency loss[J]. Sensors, 2022, 22(14): 5141.
[43] Carrara F, Amato G, Brombin L, et al. Combining GANs and autoencoders for efficient anomaly detection[C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE, 2021: 3939-3946.