

## GAN 模型生成图像检测方法综述

谢天圻, 吴媛媛, 敬 超, 孙伟恒

成都理工大学 计算机与网络安全学院(示范性软件学院), 成都 610059

**摘 要:**生成对抗网络(generative adversarial network, GAN)作为生成高质量图像的强大工具,近年来在图像合成领域得到了广泛应用。然而,随着 GAN 技术的快速发展,引发了图像伪造和欺诈的严重担忧,特别是在新闻报道、身份认证以及司法取证等关键领域。这些伪造图像不仅难以辨别,还可能被用于传播虚假信息、实施诈骗,甚至在法律案件中造成难以弥补的损害。为应对这一挑战,研究者们提出了多种检测 GAN 生成图像的方法,主要可以分为基于特征识别的方法和基于数据驱动的方法。对于当前主要的 GAN 图像检测方法进行了系统梳理,并通过重训练实验验证了它们在不同数据集上的检测准确率。对未来 GAN 图像检测领域的发展趋势进行了展望,提出了潜在的研究方向,以推动该领域的进一步创新和发展。

**关键词:**生成对抗网络(GAN);深度学习;伪造检测

**文献标志码:**A **中图分类号:**TP391 **doi:**10.3778/j.issn.1002-8331.2405-0346

### Survey of Image Detection Methods Generated by GAN Models

XIE Tianqi, WU Yuanyuan, JING Chao, SUN Weiheng

College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu 610059, China

**Abstract:** As a powerful tool for generating high-quality images, generative adversarial network (GAN) has been widely used in the field of image synthesis in recent years. However, with the rapid development of GAN technology, it also raises serious concerns about image forgery and fraud, especially in key areas such as news reporting, identity authentication and judicial forensics. These fake images not only are difficult to identify, but also may be used to spread false information, commit fraud, or even cause irreparable damage in legal cases. To cope with this challenge, researchers have proposed a variety of methods for detecting GAN-generated images, which can be mainly divided into feature-based methods and data-driven methods. This paper systematically sorts out the current main GAN image detection methods, and verifies their detection accuracy on different datasets through re-training experiments. Finally, the development trend of GAN image detection in the future is prospected, and potential research directions are proposed, in order to promote further innovation and development in this field.

**Key words:** generative adversarial network (GAN); deep learning; forgery detection

随着社交媒体的快速发展,虚假信息的传播成为现代社会关注的主要问题之一。利用各种合成软件和开源平台,人们可以轻松制作出高度逼真的伪造图像,使得图像真假难辨。这些图像在新闻报道、身份认证、司法鉴定等领域的使用,会引发严重的社会问题。例如,计算机生成的图像可以用于伪造新闻,影响社会舆论和稳定;利用合成的人脸图像伪造个人信息,不法分子可以轻松实施网络诈骗等犯罪活动;在司法鉴定中,若错误地将合成图像视为真实图像,可能会对当事人造成不

可挽回的损失。

数字图像的真实性和原始性面临前所未有的挑战。从早期发展到现在已经成熟的 GAN(generative adversarial network)模型<sup>[1]</sup>,如 BigGAN、StyleGAN 等,再到目前新兴的 Diffusion 模型,如 Stable Diffusion,这些技术不仅可以修改现有图像,还能生成原本不存在的图像。尽管这些技术的发展在某些方面带来了便利,但也无形中推动了虚假信息的传播。因此,研究和检测这类生成图像变得至关重要。

**基金项目:**成都理工大学 2023 年中青年骨干教师发展资助计划(10912-JXGG2023-06470)。

**作者简介:**谢天圻(1999—),男,硕士研究生,研究方向为图像处理、视频异常检测;吴媛媛(1986—),通信作者,女,博士,副教授,CCF 会员,研究方向为视频图像处理、计算机视觉, E-mail: wuyuanyuan@cdut.edu.cn;敬超(2000—),男,硕士研究生,研究方向为深度学习、群体情绪、视频图像处理;孙伟恒(1994—),男,工学博士,讲师,研究方向为计算机视觉、视频编码。

**收稿日期:**2024-05-24 **修回日期:**2024-07-16 **文章编号:**1002-8331(2024)22-0074-13

目前主流的检测方法可以分为两大类:基于特征检测和数据驱动型检测。基于特征检测主要通过寻找生成图像中的特征,或者学习只有真实图像才包含的特征来进行分类,这类方法只需要对特定特征进行学习,因此训练较为简单,耗时短,但往往对检测的图像具有一定针对性,即针对特定的某一类图的检测效果比较好,而泛用能力较差。数据驱动型检测则需要大量数据的堆砌,训练量往往需要数十万张图片,训练更加费时,但具有更好的泛化能力,对各种模型的生成图片都有一定的检测能力。

## 1 生成对抗网络

生成对抗网络,包括一个生成器和一个判别器,生成器学习生成能够骗过判别器的图片,而判别器则要判断出输入图片中哪些是生成器生成的假图片,二者的训练方式类似于博弈游戏,可以在不断优化下,最终达到纳什平衡。

该网络的目标函数如式(1)所示:

$$\min_G \max_D V(G,D) = \min_G \max_D E_{x \sim P_{\text{data}}} [\ln D(x)] + E_{z \sim P_z} [\ln(1 - D(G(z)))] \quad (1)$$

式中, $z$ 是服从高斯分布的随机噪声; $G$ 代表生成器; $D$ 代表判别器; $P_{\text{data}}(x)$ 表示真实数据的概率分布。

在理论证明方面,当生成器固定时,对 $V(G,D)$ 求导,可以得出最优判别器 $D'(x)$ :

$$D'(x) = \frac{P_G(x)}{P_G(x) + P_{\text{data}}(x)} \quad (2)$$

式中, $P_G(x)$ 代表生成器构造的概率分布。把最优的判别器代入目标函数,可以得到生成器 $G$ 的目标函数等价于优化 $P_{\text{data}}(x)$ 、 $P_G(x)$ 的JS散度(Jenson Shannon divergence, JSD)。当生成器和判别器的训练次数足够时,模型最终会收敛,两者达到纳什平衡。此时判别器 $D$ 对样本的判别概率均为0.5,也就是说生成的样本达到了难以区分的程度。

但是传统的GAN生成图片完全无法被控制,因此又诞生了条件生成对抗网络(conditional generative adversarial network, CGAN)<sup>[2]</sup>。在生成器的输入中,除了随机噪声 $z$ ,再加入条件变量 $c$ 用于控制生成器 $G$ 的生成结果,在判别时,除了图片也同样输入相同的条件变量 $c$ ,整体结构如图1。因此,CGAN可以获得更快的收敛速度,并且可以控制输出结果。

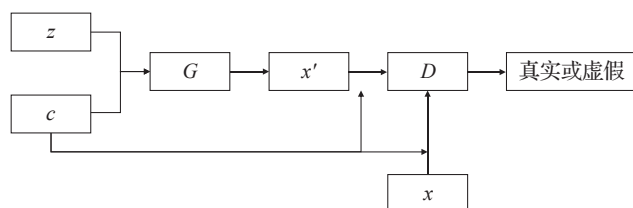


图1 CGAN结构

Fig.1 Architecture of CGAN

之后又出现了深度卷积生成对抗网络(deep convolutional generative adversarial network, DCGAN)<sup>[3]</sup>,为了使GAN能够适应卷积神经网络,DCGAN提出了以下四点架构设计:

- (1)使用卷积来替代模型中的池化层;
- (2)删除模型中的全连接层;
- (3)生成网络的激活函数改为ReLU,仅在最后输出层使用Tanh;
- (4)在生成模型和判别模型时都使用批量归一化(batch normalization, BN)。

基于以上几点优化后DCGAN模型的训练状态更加稳定,并且可以有效实现高质量图像的生成,相关生成模型在实际工程中有广泛的应用,此后的大量GAN模型都是在它的基础上进行改进。

此后谷歌的研究人员提出了BigGAN<sup>[4]</sup>,这是GAN首次生成具有高保真度和低品种差距的图像。在此之前的最高Inception得分为52.5,而BigGAN为166.3。除此之外,BigGAN还将Frechet初始距离(描述生成图像的多样性和真实性)得分从18.65提高到9.6,这些都令人印象深刻。Nvidia提出了StyleGAN<sup>[5]</sup>,该模型是一种强大的图像生成模型,以高质量的生成能力、可控性和广泛的应用领域而闻名。它在计算机视觉和人工智能领域有着广泛的影响,并为许多创新性的应用提供了潜在的解决方案。

近年来,随着模型参数的扩展和跨模态预训练模型的快速发展,基于GAN的图像生成模型在性能上取得了显著提升。Tao等人<sup>[6]</sup>实现了将CLIP(contrastive language-image pre-training)<sup>[7]</sup>嵌入到GAN模型的生成器与鉴别器中,利用了CLIP领域泛化能力引导图像生成,从而有效减少了模型训练参数和训练数据量。得益于大模型的快速发展,ChatGPT等高性能模型不断出现,也有不少研究者提出在训练GAN模型时扩大模型规模和使用更大的训练数据集。例如,GigaGAN模型<sup>[8]</sup>通过将模型参数扩展到10亿级别,并结合拥有20亿幅图像的LAION-2B<sup>[9]</sup>庞大数据集进行训练,使其能够生成性能与最先进的自回归模型或基于扩散概率模型相媲美方法。通过这些创新和改进,基于GAN的图像生成技术不断突破性能瓶颈,展现出强大的潜力和广泛的应用前景。

此外,GAN模型也有许多其他方面的应用,例如腾讯平台与内容事业群(platform and content group, PCG)开发了GFPGAN(pretrained face GAN)<sup>[10]</sup>,用于对老旧或破损的人脸照片进行还原,在还原精度和人脸的细节部分都十分优秀。另外GAN模型在视频超分辨率方面也取得了长足的进展,例如SRGAN(image super-resolution GAN)<sup>[11]</sup>,可以将低分辨率的视频进行高分辨生成,从而提高视频的清晰度。总之,GAN模型在各种图像相关领域已经有了非常丰富的应用场景。

## 2 深度学习检测方法

基于GAN架构的生成模型已经具备生成逼真图像的能力,这可能会引发安全问题。这些虚假图像可能被用于欺骗他人。例如通过在社交媒体上发布虚假新闻和伪造信息,引发各种社会问题。随着GAN技术的不断进步,虚假图像的质量也逐渐提高,可能导致更严重的后果。因此,设计能够区分真实图像和生成的虚假图像的方法变得至关重要。

尽管GAN生成的图像可以成功欺骗个体,但它们无法躲过基于计算机的人工智能驱动的检测器。接下来,将讨论基于深度学习检测虚假图像的方法。

目前的检测方法主要可以分为两类:基于特征识别(artifacts detection)和数据驱动方法(data-driven approach)。检测人工特征主要考虑在生成图像中上采样分量所留下的空间域或频率域中的周期信号,这类方法往往对于质量不高的生成图像很有作用,但对于高质量的生成图像却效果不佳。数据驱动方法学习了大量的真假图像,使得分类器可以学习生成图像中的共性特征,但是这种方法往往对算力要求较高,因此也限制了其在实际问题中的广泛应用。

### 2.1 基于特征识别

#### 2.1.1 共现矩阵

在早期的图像隐写检测中,共现矩阵方法常被用于图像分析,以识别隐藏数据。通常,将图像转换为灰度图以保留其纹理信息。共现矩阵作为特征是利用了像素对之间的相关性来捕捉图像的纹理信息。它们基于像素值在空间上的分布模式,能够有效地区分出生成图像和真实图像之间的纹理差异。

文献[12]认为图像内容无法描述局部特征,因此应该转而对残差图像进行建模并将残差特征组合,残差图像具有更窄的动态范围,可以使用共现矩阵进行更好的描述。

文献[13]提出了一种在图像RGB通道上计算共现矩阵的方法。不同于文献[12]对图像残差计算共现矩

阵,该方法直接计算红、绿、蓝三个通道上的共现矩阵,并将其输入卷积神经网络进行学习,使模型能够捕捉共现矩阵中的重要特征。

但是共现矩阵计算复杂度较高且对图像变换不敏感,因此共现矩阵这一研究方向目前暂无新的进展。

#### 2.1.2 色彩分量下的特征

文献[14]发现生成图像和真实图像在HSV和YCbCr色彩空间的色度分量上的统计特性是不同的,而在残差域的差异更为明显,于是提出由多个颜色分量的残差图像中提取的共现矩阵组成特征。该特征集具有低维数的特点,并且即使在较小的训练集上也能取得较好的检测性能。具体结构如图2所示。

该方法首先提取不同颜色空间色度分量的高频残差,然后计算基于共现矩阵的检测特征,并结合集成分类器得到最终检测结果。该方法优势主要在基于小样本数据集时,依旧可以取得较高的准确度,但从最终实验结果来看该特征并不具有普遍性,因此泛化能力并不高。

#### 2.1.3 视觉伪影

基于现有的生成方法会产生特殊的人为视觉特征,通过检测眼睛、牙齿、脸部轮廓可以简单地发现这些特征。文献[15]发现并总结了这些特征,并且将这些人为视觉特征分为以下几类:

(1)全局一致性(global consistency):包括左右眼虹膜颜色不一致。在现实中虹膜异色症相当稀少,但是这种现象在生成人脸中出现概率却很大。

(2)光照估计(illumination estimation):在用Face2face的方法中,几何估计、光照估计和渲染是显式建模的,但用深度学习生成的人脸中,对入射光的错误估计同样容易导致相关的伪影。这种伪影通常出现在鼻子周围的区域,比如一侧渲染得过黑。同样地,人眼中的反射光在生成时常常会缺失或简化为小白点,如图3所示。

(3)几何估计(geometry estimation):对人脸不精确的几何估计导致明显的边界和高对比度的人工伪影出现在蒙版和人脸的边界上。此外,部分遮挡面部的部

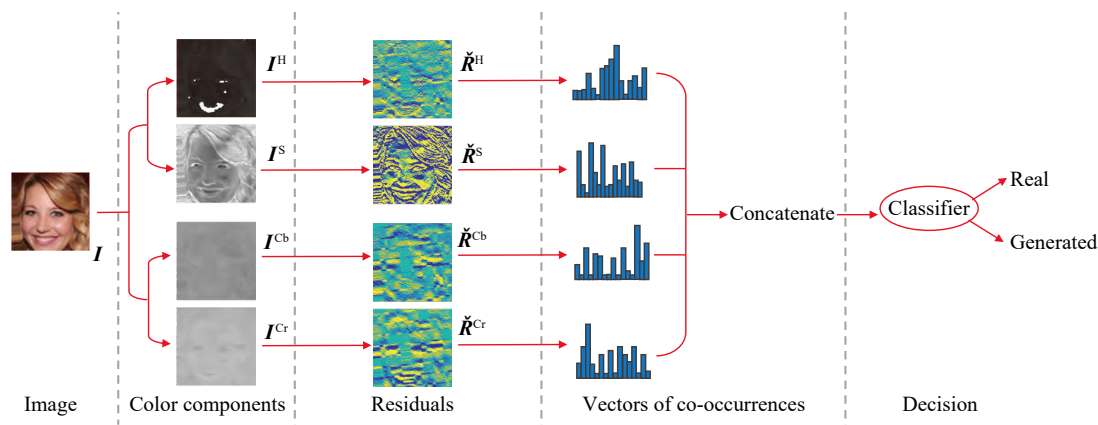
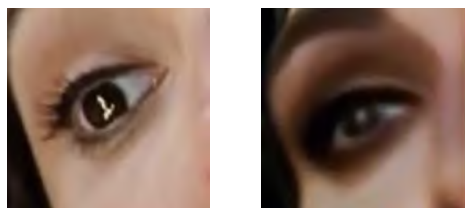


图2 算法架构图

Fig.2 Algorithm architecture





(a)真实图 (b)生成图

图3 人眼中的反射光对比

Fig.3 Contrast of reflected light in human eye

分,如头发的建模不正确,会导致出现“空洞”,如图4所示。而且牙齿通常根本没有建模。

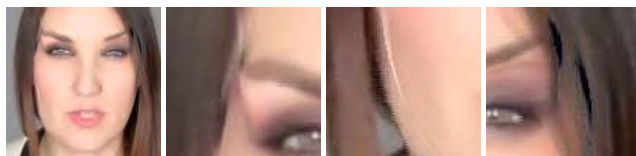


图4 头发中的“空洞”

Fig.4 “Holes” in hair

针对以上视觉特征进行检测,这种检测方法的一大优势在于检测结果的可解释性高,但同时局限性也很大,对于图像的要求较严格,例如图像中必须包含完整的人脸部分,必须有睁开的眼睛等条件,实际应用中往往无法满足要求。

#### 2.1.4 饱和度异常

文献[16]探讨了GAN生成图像的原理。通常在GAN生成中,为了结果收敛,会对数据进行归一化处理,使其在单位间隔内均匀分布,再通过“K-by-1”卷积转化为RGB值。然而,相机拍摄的图片中,像素值不受约束,入射光强常呈对数分布,需进行高动态范围成像(high dynamic range, HDR)以保留阴影和高光部分的细节。在没有HDR的情况下,照片中通常有饱和或曝光不足的区域。如图5所示,可以看到由GAN生成图像的灰度直方图基本集中在中部,两侧没有过曝或暗部区域。



(a)真实图1 (b)生成图 (c)真实图2

图5 真实图像与生成图像的灰度直方图对比

Fig.5 Comparison of grayscale histograms of real and generated images

基于上述分析,提出了一种统计图像中过曝区域和阴影区域像素频率的新颖特征提取方法,并将其应用于

支持向量机(support vector machine, SVM)模型中进行分类学习。尽管该方法在检测准确率方面表现较差,但从生成原理角度出发,提出了独特的特征提取视角,为相关领域的研究提供了新的思路。

#### 2.1.5 频率域特征

部分研究者发现,通过频域特征进行识别,不仅能够保持较高的检测准确率,还可以显著提高检测器的抗干扰能力,对常见的图像扰动表现出较强的鲁棒性。于是展开了对频域特征的研究。

##### (1) AutoGAN 模拟生成模型

文献[17]重新审视了GAN上采样模块引入的伪影,并通过信号处理分析,展示了这些伪影在频域中的表现。基于此,提出了一种基于频谱输入而非像素输入的分类器设计方法。

具体而言,对于每张图像,分别对RGB三个通道进行二维离散傅里叶变换(two-dimensional discrete Fourier transform, 2D DFT),从而获得三个通道的频谱。然后对这些频谱取对数并进行归一化处理,使其值范围在 $[-1, 1]$ 之间。经过处理的频谱作为分类器的输入,分类器通过学习图像在频域中的特征来实现对GAN生成图像的检测。

同时,提出了一种GAN模拟器 AutoGAN,能够模拟多个流行GAN模型通用的生成过程。该模拟器无需访问实际GAN模型即可生成模拟的“假”图像,用于训练分类器,有效解决了实际训练时缺少数据的问题。

##### (2) 频率域噪声模式不同

文献[18]已经证明了在GAN模型生成图像的过程中,上采样操作会导致生成的图像中出现伪影。然而,在高质量GAN模型生成的图像中,这些伪影不明显,并且大量的语义信息干扰了分类器的性能。为了解决这一问题,文献[19]提出了去噪网络,首先对图像进行去噪,然后用原始图像减去去噪后的图像,得到不受语义信息干扰的噪声模式,即学习噪声模式(learned noise patterns, LNP)。

对图像进行二维离散傅里叶变换后,可以在振幅谱中看到生成图像存在不同周期的震动模式。如图6所示,StyleGAN和StyleGAN2<sup>[20]</sup>采用双线性插值进行上采样,其LNP振幅谱的周期为8;CycleGAN使用反卷积上采样,LNP振幅谱的周期为4。生成图像在振幅谱中表现为网格状伪影,而真实图像的LNP在频域上非常相似。因此,提出通过学习真实图像的LNP来区分生成图像。

总体而言,该文提出的检测方法通过利用真实图像固有的频域特性,实现了对生成图像的高效检测,为生成图像的鉴定提供了新的视角和强有力的工具。

网络结构如图7所示,首先输入图像经过LNP提取模块,该模块包含4个递归残差块(recursive residual group, RRG)<sup>[21]</sup>用于提取出LNP振幅谱,然后输入到特征融合模块中,最后输出结果进行分类判别。

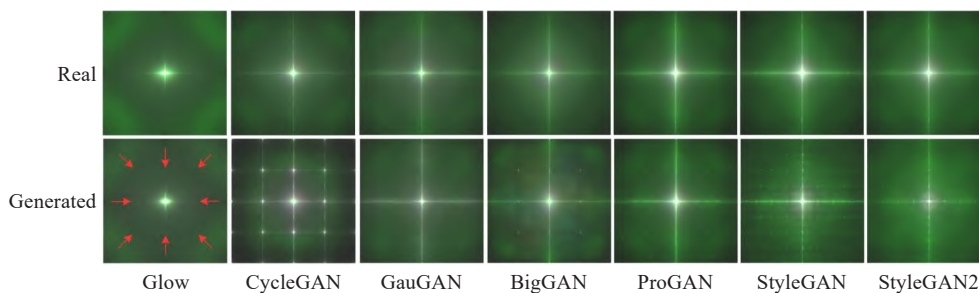


图6 真实图像与生成图像振幅谱对比

Fig.6 Comparison of amplitude spectra between real and generated images

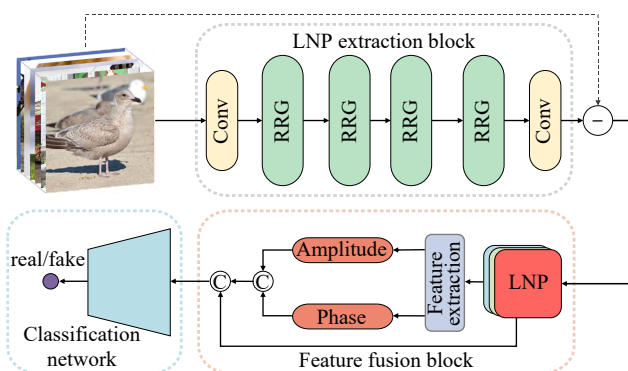


图7 文献[19]中模型结构图

Fig.7 Model structure in reference [19]

### (3) 频率域共有的伪影

与前文不同的是,文献[22]使用离散余弦变换(discrete cosine transform, DCT)将图像变换到频率域。生成图像经过变换后与真实图像相比,生成图像的光谱总会在上部、左部产生高频值,如图8所示。

经过频域特征提取后,使用简单的线性分类器即可实现高准确度的生成图像检测,表明了频域特征在区分生成图像和真实图像方面的显著优势。与其他方法相比,该研究发现频域分类器对抗常见的图像扰动更具有鲁棒性。这一发现表明,频域特征不仅能提高分类器的准确性,还能增强其在实际应用中的可靠性。

同时该方法也存在一些缺陷:首先,该方法依赖于GAN生成图像中的频域伪影,如果未来的GAN模型改进上采样技术,减少或消除这些伪影,方法的有效性可能下降。其次,处理高分辨率图像时,该方法面临较高的计算资源需求,增加了处理成本和复杂性。尽管频域分类器在实验中表现良好,但在面对未见过的GAN模

型时,其泛化能力有限,需要重新训练以适应新的数据分布。使用频域分析需要复杂的数据预处理,这对计算环境和硬件有较高要求,限制了其在资源有限环境中的应用。

### 2.1.6 局部与整体特征

对于图像整体进行检测时,检测器常常会受到图像语义信息的干扰,导致无法准确学习生成图像的特征。因此,部分研究者提出采用图像的局部信息作为特征提取区域进行识别。

#### (1) 基于图像块进行识别

文献[23]提出了一种基于图像块(patch)的分类器,限制模型的感受野,使其专注于图像的小区域而非全局结构。这种方法能够更好地捕捉局部伪造特征,从而增强分类器的检测能力。此外,通过对图像进行patch级别的分类预测,可以生成热图,展示哪些区域更容易被检测为伪造。这种热图有助于理解伪造图像中最具指示性的特征,从而提高检测器的可解释性。但该方法仅仅针对人脸图像进行检测,对于其他种类的生成图像无法进行有效检测。图9展示了该方法的简要结构以及生成的热图结果。

#### (2) 融合整体和局部特征

局部与整体特征同时考虑图像的局部细节和全局结构。局部特征侧重于分析图像中特定区域的纹理信息或者局部像素的统计特性,而整体特征则关注整幅图像的整体分布模式和统计特征。这种综合分析有助于检测器更全面地理解图像的结构和内容。

文献[24]设计了一个双分支模型,包括全局分支和局部分支,通过结合全局和局部特征来增强检测能力。

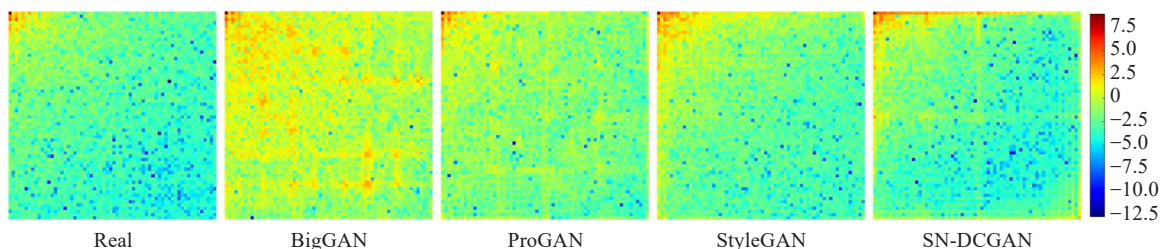


图8 离散余弦变换后图像的频域对比

Fig.8 Frequency domain comparison of images after discrete cosine transform



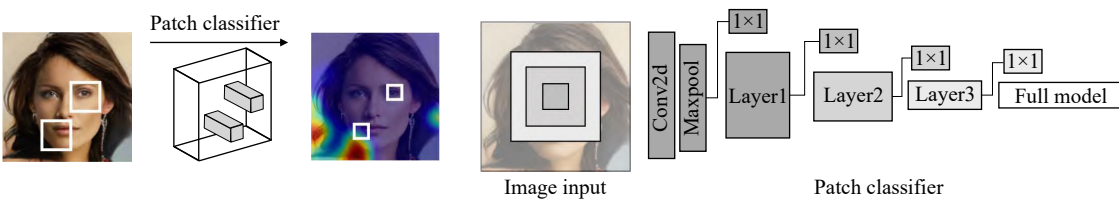


图9 文献[23]中模型简要结构

Fig.9 Brief structure of model in reference [23]

全局分支负责从整个图像中提取全局空间信息,局部分支则从多个选定的局部 patch 中提取细粒度特征。为了高效融合全局和局部特征,文章设计了一个基于多头注意力机制的特征融合模块(attention-based feature fusion module, AFFM)。通过多头注意力机制计算所有嵌入之间的相关性,并生成一个融合后的特征向量,该向量用于最终的二分类预测。

该双分支模型结构如图 10 所示,通过结合全局和局部特征,并利用注意力机制进行特征融合,该方法显著提高了生成图像检测的准确性和鲁棒性,为生成图像检测领域提供了一种有效的解决方案。

尽管这篇文章的方法表现出色,但也存在一些局限性。首先,patch选择模块(patch selection module,PSM)依赖于全局特征图,如果全局特征图质量不高,可能影响检测性能。此外,处理全图和多个 patch 时需要较高的计算资源和时间,尤其在高分辨率图像情况下。

(3) 贫富纹理区域特征

Zhong 等人<sup>[25]</sup>提出了一种对比纹理丰富区域与纹理贫乏区域像素相关性的方法。由于纹理丰富区域的分布比纹理贫乏区域更为复杂,生成模型从随机噪声中合成纹理丰富区域的难度更大。此外,作者还指出了另一

关键点——像素相关性(类似于光响应不均匀性(photo response non-uniformity, PRNU)噪声模式<sup>[26-27]</sup>),其主要取决于拍摄图像的 CMOS(complementary metal oxide semiconductor)和 ISP(image signal processing),因此,真实图像的纹理丰富区和纹理贫乏区具有很高的像素相关性。作者利用这一点作为主要判断依据来识别生成图像。这种更为通用的特征使得该方法具有良好的泛化能力,在检测 GAN 模型生成图和 Diffusion 生成图时均表现出较高的准确率。

模型的主要结构如图 11 所示,该方法可以分为图像特征提取和分类器训练两部分。为了抑制图像的语义信息并提取像素间的相关性,首先对图像进行随机裁剪,并根据裁剪块的纹理丰富性进行排序,将裁剪后的块分为纹理丰富区和纹理贫乏区。纹理的多样性可以通过像素波动程度来衡量。此外,还使用了 SRM(spatial rich model)<sup>[28]</sup>提出的一组高通滤波器来提取这两个区域的噪声模式,通过高通滤波器可以有效抑制图像语义信息并放大像素间的相关性,用于识别各种伪造图像。最终,通过卷积层和归一化处理后进行分类。该方法的一个优点是对输入图像的大小没有特殊要求,特征提取仅与裁剪的像素块大小有关,因此无论输入图像过大或

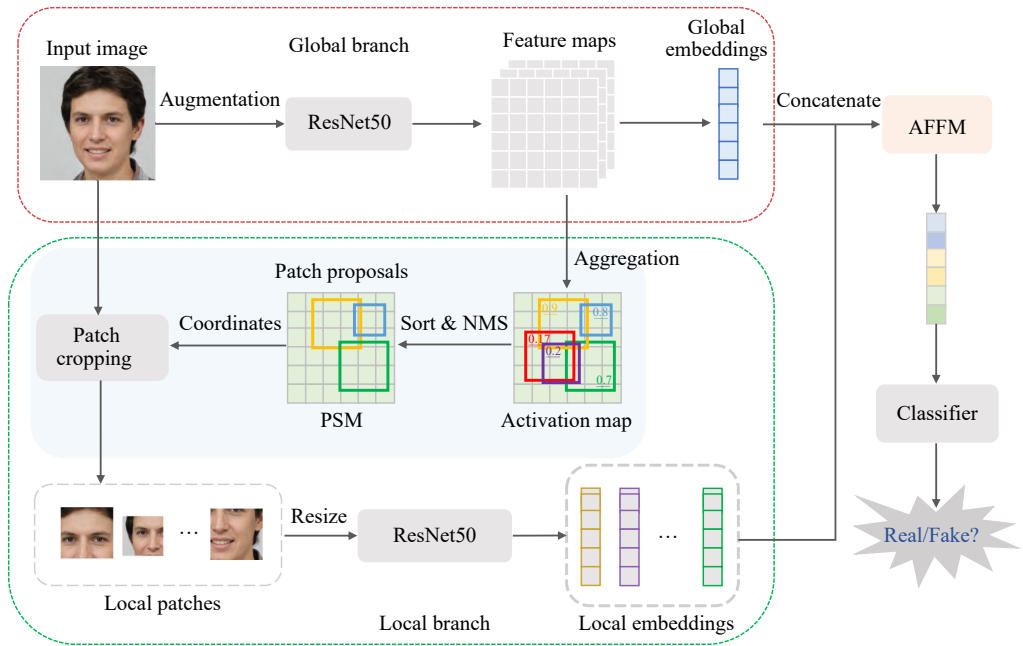


图10 文献[24]中模型整体结构图

Fig.10 Overall structure of model in reference [24]

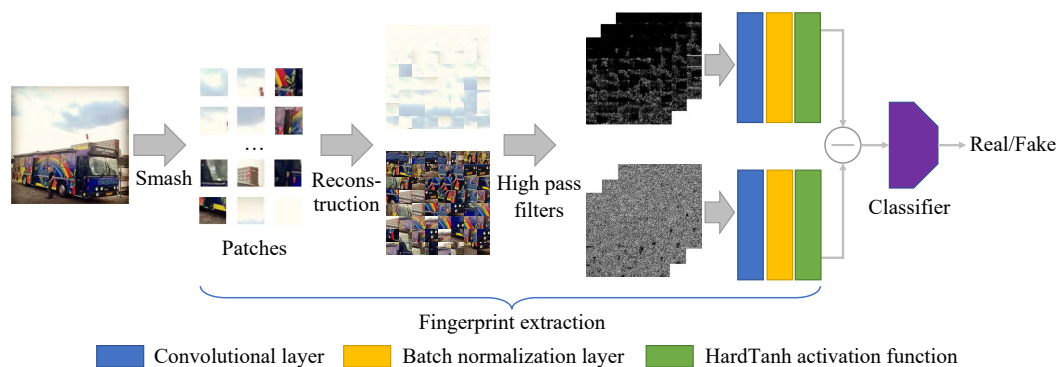


图 11 文献[25]中模型整体结构图

Fig.11 Overall structure of model in reference [25]

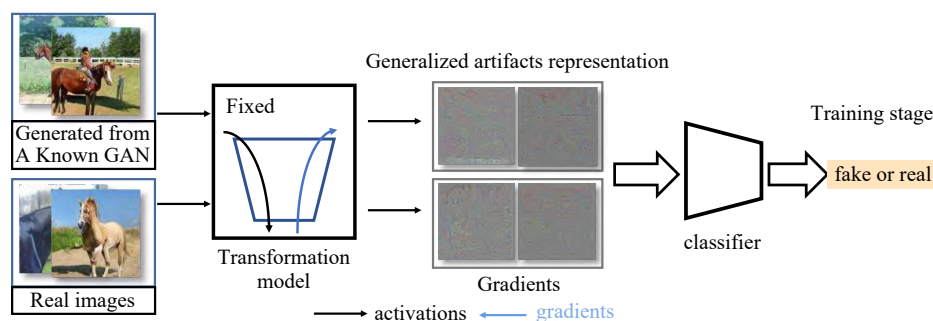


图 12 文献[29]中模型训练结构图

Fig.12 Model training structure in reference [29]

过小,均可以进行相同的处理,不会因图像缩放处理而导致模型性能下降。

### 2.1.7 梯度特征

梯度特征基于像素值变化的强度来区分生成图像和真实图像。这种方法不仅可以用于边缘检测,还能作为生成图像检测的有效工具之一,因为生成图像通常在梯度分布上表现出与真实图像不同的模式。

#### (1) 梯度学习

文献[29]提出了梯度学习(learning on gradients, LGrad)方法,首次利用预训练卷积神经网络(convolutional neural network, CNN)的梯度作为GAN生成图像特征的表示。预训练CNN模型的梯度是损失函数相对于输入图像的导数,表示每个像素对模型输出的影响。这些梯度过滤掉了图像内容,仅保留与预训练模型目标任务相关的判别像素。使用这种方法,有效地解决了GAN生成图像检测中的数据依赖问题,显著提高了检测性能和鲁棒性。

该文设计了一种新的检测框架,以提高生成图像检测的泛化能力。如图12所示,该框架采用梯度作为生成图像的特征,从而获得更鲁棒的检测性能。通过使用预训练的CNN模型,将图像转换为梯度表示。在转换过程中,由于CNN的池化层,图像中大部分内容会被过滤掉,仅保留对转换模型至关重要的像素。

如图13所示,转换后的结果中,灰色部分展示了梯度的可视化效果,而绿色部分显示了在红、绿、蓝分量上

的热图,其中红色表示高值区域。在生成的梯度图中,图像的具体内容被抹去,仅保留了用于预训练模型分类任务的判别像素,利用这种梯度作为生成图像的特征来构建新的检测框架。使用梯度进行分类的主要原因是,过去的检测器往往过于依赖训练数据,导致对于未见过的图像难以进行准确识别。

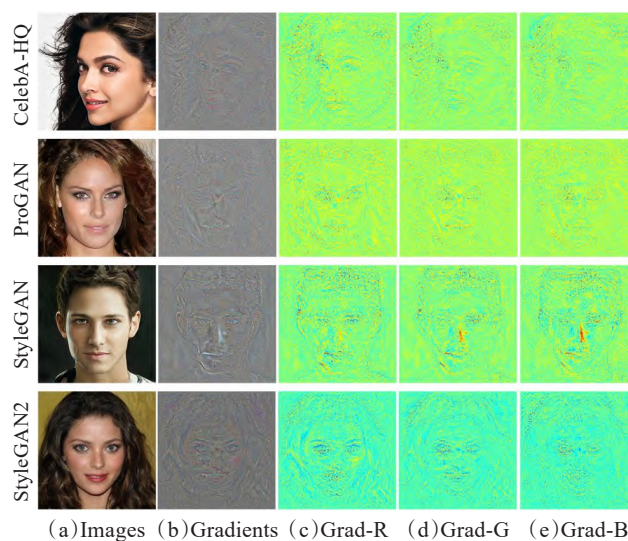


图 13 文献[29]中梯度可视化图以及红绿蓝三通道的热力图

Fig.13 Gradient visualization in reference [29] and heat maps of red, green and blue color channels

然而,该方法对于非GAN模型生成的图片(如deepfake)的检测效果不佳。同时文章中提到的方法依

依赖于预训练的CNN模型来提取图像的梯度,这些梯度作为伪影的通用表示被用于GAN生成图像的检测。虽然这种方法在特定预训练模型上表现良好,但其有效性在不同类型的预训练模型上可能会有所差异。尤其是在实际应用中,不同的预训练模型可能会对不同类型的GAN生成图像产生不同的响应,导致检测的鲁棒性和泛化能力下降。

## (2)生成图片归类

文献[30]首次系统地研究了GAN生成图像中存在独特且稳定的指纹特征,这些指纹可以用于区分不同GAN模型生成的图像。GAN模型指纹是指不同GAN模型在生成图像过程中,由于训练数据、架构和训练过程的不同,会在生成的图像中留下特定的模式和伪影。对于每个GAN模型,这些模式和伪影即为模型的指纹。通过实验证明了GAN指纹在不同图像频率和图像块大小中保持稳定,不受伪影的影响。通过对指纹的微调,展示了这些指纹对多种对抗图像扰动的抵抗力。

该研究不仅在理论上证明了GAN指纹的存在和独特性,还通过实验验证了这些指纹在实际应用中的可行性和有效性,提供了一个可以直接应用于视觉取证和知识产权保护的工具。

简要结构如图14,训练时将每个图像与其来源组成图-源对,学习图像指纹。与自然语言处理中的作者归属任务类似,其训练了一个归属分类器,该分类器的目标就是预测图片的来源究竟是真实图像还是某种GAN模型。

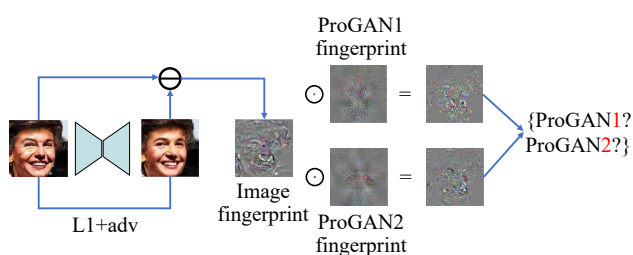


图14 文献[30]中模型的简要结构图

Fig.14 Brief structure of model in reference [30]

但这种方法也存在一些缺陷:由于该方法依赖大量不同GAN模型生成的图像作为训练数据,若训练数据集中不包含某些特定的GAN模型或伪影,分类器可能无法有效识别这些未见过的图像。此外,分类器可能会在训练数据上过拟合,从而在测试时表现不佳,特别是在测试数据分布与训练数据有较大差异时。虽然文中提供了指纹的可视化方法,但这些可视化结果对于非专家用户来说可能仍然难以理解。如何将指纹可视化结果转化为直观的解释仍然是一个挑战。

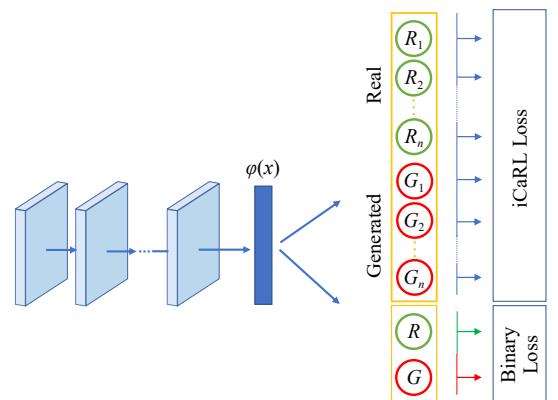
## 2.2 数据驱动方法

从上一节的讨论中可以得出,训练一个深度神经网络来对真实和生成图像进行分类已经不再是主要难题,

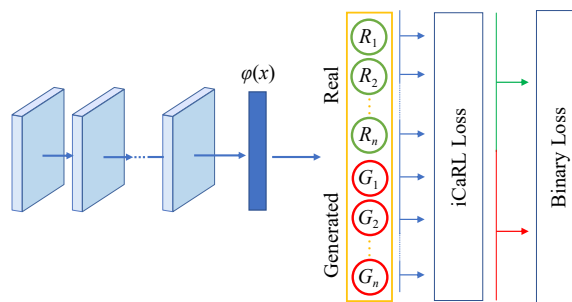
更加令人关注的是如何将检测器推广到识别未知GAN模型生成的图像。

文献[31]提出,通过增量学习的方法在检测器的训练中不断增加新的GAN模型数据进行学习,从而提高模型的泛化能力,以应对不断更新的GAN模型生成的图像。关于增量分类方法的研究文献非常丰富,这种方法能够在无需重新训练整个数据集的情况下,适应越来越多类别的分类任务<sup>[32-33]</sup>。与传统场景不同的是,在训练时,该方法会训练来自所有类别的所有数据,而在类增量场景中,随着时间的推移,新的类的数据会随着训练进行而不断出现。因此,需要一种特殊的类增量学习策略,使得在新种类数据集不断出现时有效地更新分类方法。这种策略成功的关键在于避免在适应新训练数据的同时忘记以前的分类能力,即解决灾难性遗忘的问题。但有效解决此问题后,训练出来的模型将具有很好的泛化能力,可以适用于绝大部分的检测任务中,大大减少了重新训练整个模型的时间和资源消耗。

文中选择了一种有效的增量学习算法iCaRL(incremental classifier and representation learning)<sup>[34]</sup>,在此算法基础上提出两种方法,具体结构如图15所示。图15(a)为多任务多分类器(multi-task multi-classifier, MT-MC),在该方法中,检测器与增量分类器并行运行;图15(b)为多任务单分类器(multi-task single classifier, MT-SC),在分类的同时也进行检测。以上算法都基于通用的CNN的分类器和从所有可用类中均匀提取一组示例图像 $P$ 。值得注意的是,CNN的训练和更新只是为了作为特征



(a)多任务多分类器(MT-MC)



(b)多任务单分类器(MT-SC)

图15 文献[31]中提出的方法结构图

Fig.15 Method structure proposed in reference [31]



提取器,而实际的分类是基于与范例图像相关的特征向量进行的。范例图像本身是避免忘记以前知识的关键。

增量学习在面对大规模数据时,必须提供新GAN模型的架构,这导致在实际使用中的实用性大大降低,同时所花费的时间成本也较高。但是从另一方面来说,对比其他泛化能力更弱的方法,增量学习可以在不重新训练整个模型的情况下迅速适应新数据,在技术飞速发展的今天也具有一定的优势。

文献[35]证明,一个标准的深度卷积神经网络在单个生成器的数据上训练后,依旧可以在未知GAN模型生成的图像上保持良好的检测效率。在LSUN(large-scale scene understanding)的20个类别共72万张图片上进行训练,使用的网络为ResNet-50<sup>[36]</sup>,训练时仅使用ProGAN的生成图像,并在训练前对图片进行四组预处理:无增广、高斯模糊、JPEG压缩+高斯模糊、JPEG压缩。这些预处理的最终结果能够有效区分各种生成模型生成的图像和真实图像。

该结果进一步表明,通过随机翻转、模糊和JPEG压缩等数据增强操作后,训练出的检测器具有更高的泛化性。由于高斯模糊和图像压缩,可以强迫模型学习更加通用的特征,泛化性十分优秀。但是最新的研究证明,使用该方法在GAN生成图像数据集预训练的模型,无法准确检测目前更加热门的Diffusion模型的生成数据<sup>[37]</sup>。

在这一基础上,文献[38]提出no-down模式,即去除第一层中的下采样步骤,这一改进可以将平均准确率从80.71%提高到94.42%,但同时也增加了更多的可训练参数,显著延长了训练时间。

## 2.3 小结

在生成对抗网络模型生成图像检测方法的研究中,研究者们探索了多种不同的特征和算法,以提高对生成图像的检测能力。本节介绍了各种检测方法,包括共现矩阵特征法、频率域特征法、局部特征和整体特征法等。每种方法都有其独特的优势和适用场景。表1对这些检测方法进行了简要对比和总结。

表1 生成图像检测方法对比

Table 1 Comparison of generated image detection methods

类别	名称	方法	特点	优势与不足
基于特征识别	共生矩阵	文献[13]	提取图像的RGB三个色彩通道上的共现矩阵作为特征进行识别	网络结构简单,训练快捷;但泛用性差,仅在训练模型中有较高识别率
	色彩分量	文献[14]	提取HSV和YCbCr色彩空间的共现矩阵作为特征	集成学习方法准确率高;但训练较为复杂,面对未知模型时性能会明显下降
	视觉伪影	文献[15]	人脸的眼睛、牙齿、脸部轮廓作为判断特征	特征较为简单,可解释性强;但随着生成技术发展,该特征难以被应用
	饱和度	文献[16]	从GAN模型生成原理出发,提出像素值高频和低频出现频率作为检测特征	对于算力要求较低,特征提取较简单;但该特征目前基本没有泛用性,仅在特定GAN模型中可以使用
		文献[17]	对RGB三个通道分别进行傅里叶变换后,将这三个通道的频谱作为特征输入,并提出了AutoGAN模型来模拟不同GAN模型的生成结果	有效解决了整理训练数据的问题,准确率较高;但训练较为复杂,模型泛用性不高
	频率域	文献[19]	学习真实图像的噪声信息中的频率模式	检测准确率较高;但是对于jpeg格式的压缩图像,压缩时会产生相同的频率模式,因此无法识别
		文献[22]	通过离散余弦变换将图像变换到频率域,分析GAN模型生成图像的频率域模式发现了新的特征	特征的提取方法较为简单,训练量小;但该模型抗干扰能力较差,可以针对其提取的特征进行优化生成结果,使之无法检测
		文献[23]	训练了一个局部特征提取器,专注于提取生成图像局部的特征	模型的可解释能力较强,能够明确识别的位置;但该模型主要针对人脸图像的检测,无法稳定地检测所有种类的生成图像
	局部与整体	文献[24]	双分支模型,分别对整体和局部进行分析	检测准确率较高;但模型训练比较复杂,较为耗时
		文献[25]	提出图像贫富信息区域作为特征进行检测,将图像随机裁切成小块,选择富信息区域进行特征提取	检测效率较高;但一旦图像经过高通滤波后,检测准确率有明显下降
梯度特征		文献[29]	将图像经过CNN变换为梯度图,以此作为特征进行检测	排除了图像语义信息的干扰;但对于非GAN模型检测效果不佳
		文献[30]	提出了对生成图片进行归因的研究,即检测图像是由何种模型生成	不仅可以检测出是否为生成图像,甚至可以对图像进行溯源;但模型参数较多,需要记录每种生成模型的特点,训练较为费时
增量学习		文献[31]	不断增加新的GAN模型进行学习,从而提高检测的泛化能力	能够有效提高检测的泛化能力;但训练需要不断地增加新模型,十分费时
数据驱动	CNN+高斯模糊	文献[35]	用庞大的数据量训练,提高模型识别精度和泛化能力	识别准确率高,泛化能力强;但训练数据量十分庞大,训练较为费时
	CNN+no-down	文献[38]	在卷积的同时去除第一层的下采样步骤	增加了可训练的参数,有效提高了准确率;但训练时长大大增加

在实际应用中,检测的图片往往不是直接生成,而是通过社交平台进行传播,在这一过程中不可避免地会在图像中加入噪声或压缩。这种数据干扰可能极大地影响检测模型的效果。为了解决这一问题,可以使用从机器学习角度出发的泛化技术,例如数据增强,对训练数据进行预处理,如模糊、翻转或加入不同程度的噪声等。已有研究表明,这种数据增强处理能够有效提高检测模型的泛化能力。

此外,还可以从生成模型的基础架构出发,考虑生成图像中的共有特点。例如,乔通等人<sup>[39]</sup>提出,自然图像通常经过透镜、光学滤波、传感器光电转换、CFA插值等线性和非线性变换后形成,因此生成图像与自然图像在色彩通道上会有些许差异。文献[19]也提出,通过学习真实图像的噪声模式来进行分类,是一种单类学习方法。在负样本种类非常多且难以全面获取的情况下,单类学习方法能够有效提高模型的泛化能力。

总体来说,通过数据增强和深入理解生成模型的架构,可以显著提升检测模型在实际应用中的性能,增强应对各种干扰的能力。

2.4 对比实验

本次实验测试集来自文献[25]收集整理,包括16种不同的生成模型,包含GAN生成模型和Diffusion生成模型两大类,具体信息如表2所示,部分实验数据也来自文献[25]。为了保证实验结果的公平性,对于进行实验的模型均使用同一训练集进行重新训练,该训练集来自文献[35],包括72万张图片,其中36万张真实图片来自LSUN,36万张生成图片来自ProGAN模型。由于本任务暂时没有公开数据集,大部分使用数据真实图像来自LSUN或ImageNet,生成图像则直接由对应模型进行生成。

其中,LSUN是一个用于计算机视觉研究的大规模场景理解数据集。它提供了多个场景类别的高分辨率图像,可用于图像分类、目标检测、场景分析等任务。而ImageNet数据集是一个计算机视觉数据集,是由斯坦福大学的李飞飞教授带领创建。该数据集包括14 197 122张图片和21 841个Synset索引。

表2 实验数据集大小及来源

Table 2 Experimental dataset size and source

生成模型	大小	数量/ $10^3$	来源
ProGAN <sup>[40]</sup>	256×256	8.0	LSUN
StyleGAN <sup>[5]</sup>	256×256	12.0	LSUN
BigGAN <sup>[4]</sup>	256×256	4.0	ImageNet
CycleGAN <sup>[41]</sup>	256×256	2.6	ImageNet
StarGAN <sup>[42]</sup>	256×256	4.0	CelebA
GauGAN <sup>[43]</sup>	256×256	10.0	COCO
StyleGAN2 <sup>[20]</sup>	256×256	15.9	LSUN
WFIR <sup>[44]</sup>	1 024×1 024	2.0	FFHQ
ADM <sup>[45]</sup>	256×256	12.0	ImageNet
GLIDE <sup>[46]</sup>	256×256	12.0	ImageNet
Midjourney <sup>[47]</sup>	1 024×1 024	12.0	ImageNet
SDv1.4 <sup>[48]</sup>	512×512	12.0	ImageNet
SDv1.5 <sup>[48]</sup>	512×512	16.0	ImageNet
VQDM <sup>[49]</sup>	256×256	12.0	ImageNet
Wukong <sup>[50]</sup>	512×512	12.0	ImageNet
DALL-E <sup>[51]</sup>	256×256	2.0	ImageNet

首先, GAN模型生成图像数据集的检测结果如表3所示,对于每种数据集检测准确率最高的方法进行了加黑显示。数据集中正负样本是均匀分布的,因此对于准确率低于50%的方法进行了添加下划线显示,表示其对于该类数据集完全无法检测。从实验结果可以看出:大部分检测方法对于GAN模型的生成图像都可以进行有效检测。部分检测方法的检测平均准确率可以达到89%以上,如文献[19]、[29]、[25]方法的准确率也几乎达到了89%。这其中有一些共同点,这些方法都有意排除了图像的语义信息干扰,其中文献[19]将图像转化到频率域提取特征进行识别;文献[29]将图像转换为梯度图;文献[25]则是将图片分割成小块,并将每个图像块的噪声信息作为特征,也排除了图像语义信息的干扰,从而有效提高了检测泛化能力。但少数方法的泛用性太差,以至于只能对训练的GAN模型数据集进行有效检测,而一旦更换生成模型则无法进行检测,如文献[17],因采用线性分类器,导致模型泛化能力较差。由此可以看出,目前针对GAN模型的检测方法还有许多改进空间。首先是提高检测模型的泛用性,使其不仅仅针对某一个模型的生成图像进行检测;其次是如何在对大多数生成

表3 GAN模型生成图像数据集的检测结果

Table 3 Detection results of graph dataset generated by GAN model

单位: %

数据集	文献[35]	文献[22]	文献[24]	文献[19]	文献[29]	文献[25]	文献[13]	文献[17]	文献[14]
ProGAN	<b>100.00</b>	99.36	<b>100.00</b>	99.67	99.83	<b>100.00</b>	99.99	94.65	92.13
StyleGAN	90.17	78.02	85.20	91.75	91.08	92.77	<b>99.20</b>	<u>46.22</u>	86.25
BigGAN	71.17	81.97	77.40	77.75	85.62	<b>95.80</b>	51.83	57.26	65.35
CycleGAN	<b>87.62</b>	78.77	87.00	84.10	86.94	70.17	69.86	<u>38.79</u>	70.02
StarGAN	94.60	94.62	97.00	99.92	99.27	<b>99.97</b>	62.95	<u>32.98</u>	96.80
GauGAN	81.42	80.57	77.00	75.39	78.46	71.58	51.77	<b>95.56</b>	62.79
StyleGAN2	86.91	66.19	83.30	94.64	85.32	89.55	<b>99.00</b>	62.16	85.84
Average	87.41	82.79	86.70	89.03	<b>89.50</b>	88.55	76.37	61.09	79.88

图像检测准确率较高的情况下,对模型进行优化,使检测模型能够在已有的基础上提高对新出现的模型的检测能力。

在以上实验基础上,又进行进一步实验,测试检测器在新型模型包括 Diffusion 模型及其他新型的图片生成模型上的检测效果。实验结果如表4所示。

由于生成模型技术的不断发展,相较于 GAN 生成模型,目前 Diffusion 模型通过逐步添加噪声到图像中,并在训练过程中学习如何消除这种噪声从而生成新的图像,这种方法更好地模拟了复杂图像分布,生成图像的质量更高。因此本次实验也加入了 Diffusion 相关的数据集进行比较。从实验结果中可以看出,目前大部分的生成图像检测方法,对于 Diffusion 相关数据集都无法进行有效检测,大部分检测方法在 Diffusion 相关数据集上的表现明显不如在 GAN 相关数据集上的表现。这主要是因为这些方法通常仅在 GAN 模型生成的图像数据集上进行训练,所以难以适应 Diffusion 模型生成的图像。然而,实验中也有少数检测方法能够兼顾对 Diffusion 生成图像的检测。如文献[19]采用了训练数据中仅有正样本的单类学习法。单类学习法不依赖于负样本(即生成图像),而是通过学习正样本(即真实图像)的特征来识别异常。这种方法在训练时仅需真实图像数据,因此能够有效地检测到 GAN 模型和 Diffusion 模型生成的图像。而文献[25]主要对富纹理区域进行检测。富纹理区域通常包含丰富的细节和复杂的模式,这些区域在生成图像中往往最难被模拟真实。文献[29]主要通过集中检测这些区域的特定特征,能够较好地区分真实图像和生成图像。无论是 GAN 还是 Diffusion 模型,在这些富纹理区域都会产生一些特定的人工特征,因此该方法在对 Diffusion 模型生成图像进行检测时也表现出较高的准确率。

尽管现有的大部分检测方法在 Diffusion 模型生成图像上的表现不佳,但通过采用不同的检测策略,如单类学习和富纹理区域检测,能够显著提升检测效果。未来的研究可以从以下几个方面进行改进:

(1)结合 GAN 和 Diffusion 模型生成图像的特点,开发新的检测方法,能够在不同生成模型上均表现出色。

(2)扩展训练数据集,包含更多种类的生成模型生成的图像,提升检测模型的泛化能力。

### 3 总结与展望

目前人工智能领域已经有了很多创新,但安全风险却没有得到足够重视。基于人工智能的创新已经或可能在未来几年构成威胁,因此更应该关注目前生成模型的安全问题。

本文总结了对于 GAN 模型生成图像的检测方法,并分析了其优点和缺陷。通过实验比较了其中大部分方法的检测准确率和泛用性。为了保证实验结果的公平性,本文采用了复现并重训练的方式,使得每种方法均基于同一种数据集进行训练后,再使用测试数据集进行检测实验,总结并分析了实验结果。

从实验结果来看,目前的检测方法大多都还有改进空间。大多数检测方法缺乏泛化能力,许多基于特征检测的方法仅仅适用于特定的 GAN 模型,而缺乏对于新出现的 GAN 模型生成图像的检测能力。在图像处理技术发展如此迅速的今天,新的图像生成模型层出不穷;在 GAN 模型后又出现了 Diffusion 模型,现有的检测方法也基本无法直接对其生成的图像进行检测。

未来对于 GAN 模型检测方法的优化可以集中在以下几点:

(1)增强特征通用性:研究更加通用的特征提取方法,使检测器能够在不同的生成模型和图像条件下保持高准确率。例如,深入研究高频特征和像素相关性的提取与利用。

(2)鲁棒性提升:进一步增强检测器在图像压缩、噪声等干扰下的检测能力。开发更强大的数据增强技术,如多种压缩和噪声模型的模拟,以提升检测器的抗干扰性能。

(3)增量学习与在线学习:随着新型 GAN 模型的不断出现,检测器需要具备增量学习和在线学习的能力,

表4 Diffusion模型以及其他生成图像数据集的检测结果

Table 4 Detection results of Diffusion model and other generative graph dataset

单位:%

数据集	文献[35]	文献[22]	文献[24]	文献[19]	文献[29]	文献[25]	文献[13]	文献[17]	文献[14]
WFIR	<b>91.65</b>	50.75	66.80	70.85	55.70	85.80	<u>46.20</u>	<u>42.38</u>	61.75
ADM	60.39	63.42	<u>49.00</u>	<b>84.73</b>	67.15	82.17	74.96	<u>48.13</u>	63.76
GLIDE	58.07	54.13	57.20	<b>80.52</b>	66.11	83.79	75.80	<u>41.27</u>	70.71
Midjourney	51.39	<u>45.87</u>	52.20	65.55	65.35	<b>90.12</b>	72.89	<u>47.37</u>	66.05
SDv1.4	50.57	<u>38.79</u>	51.00	85.55	63.02	<b>95.38</b>	75.67	<u>46.22</u>	68.11
SDv1.5	50.53	<u>39.21</u>	51.40	85.67	63.67	<b>95.30</b>	75.08	<u>45.91</u>	68.82
VQDM	56.46	77.80	55.10	74.46	72.99	<b>88.91</b>	62.95	55.14	66.59
Wukong	51.03	<u>40.30</u>	51.70	82.06	59.55	<b>91.07</b>	75.43	<u>48.45</u>	64.84
DALL-E2	50.45	<u>34.70</u>	52.80	88.75	65.45	<b>96.60</b>	69.86	<u>37.16</u>	70.70
Average	57.84	<u>49.44</u>	54.13	79.79	64.33	<b>89.90</b>	69.87	<u>45.78</u>	66.81



能够在无需重新训练整个模型的情况下,快速适应新模型生成的图像。这可以通过设计更有效的增量学习策略和在线更新机制来实现。

(4)多模态检测:结合多种数据模态进行综合检测,以提高对复杂伪造内容的识别能力。例如,结合图像与其关联文本的语义信息进行跨模态检测。

(5)可解释性:提升检测器的可解释性,使其能够明确指出图像中被判定为生成图像的具体区域和特征。这不仅有助于理解检测过程,还能为用户提供更直观的判断依据。

(6)实时检测与部署:开发轻量级、高效的检测算法,以便于在实际应用中进行实时检测和部署。优化模型结构和计算效率,使检测器能够在各种设备和平台上运行,包括移动设备和嵌入式系统。

(7)开放数据集与基准测试:建立更大规模、更多样化的公开数据集和标准测试基准,推动检测技术的研究与发展。通过公开数据和基准测试,可以更好地评估和比较不同检测方法的性能,促进技术进步。

## 参考文献:

- [1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27, 2014.
- [2] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv:1411.1784, 2014.
- [3] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv:1511.06434, 2015.
- [4] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[J]. arXiv:1809.11096, 2018.
- [5] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4401-4410.
- [6] TAO M, BAO B K, TANG H, et al. GALIP: generative adversarial clips for text-to-image synthesis[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14214-14223.
- [7] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.
- [8] KANG M, ZHU J Y, ZHANG R, et al. Scaling up GANs for text-to-image synthesis[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 10124-10134.
- [9] SCHUHMANN C, BEAUMONT R, VENCU R, et al. LAION-5B: an open large-scale dataset for training next generation image-text models[C]//Advances in Neural Information Processing Systems 35, 2022: 25278-25294.
- [10] WANG X, LI Y, ZHANG H, et al. Towards real-world blind face restoration with generative facial prior[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9168-9178.
- [11] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4681-4690.
- [12] COZZOLINO D, GRAGNANIELLO D, VERDOLIVA L. Image forgery detection through residual-based local descriptors and block-matching[C]//Proceedings of the 2014 IEEE International Conference on Image Processing, 2014: 5297-5301.
- [13] NATARAJ L, MOHAMMED T M, CHANDRASEKARAN S, et al. Detecting GAN generated fake images using co-occurrence matrices[J]. arXiv:1903.06836, 2019.
- [14] LI H, LI B, TAN S, et al. Identification of deep network generated images using disparities in color components[J]. Signal Processing, 2020, 174: 107616.
- [15] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops, 2019: 83-92.
- [16] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, 2019: 4584-4588.
- [17] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in GAN fake images[C]//Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security, 2019: 1-6.
- [18] KINGMA D P, DHARIWAL P. Glow: generative flow with invertible  $1 \times 1$  convolutions[C]//Advances in Neural Information Processing Systems 31, 2018.
- [19] BI X, LIU B, YANG F, et al. Detecting generated images by real images only[J]. arXiv:2311.00962, 2023.
- [20] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8110-8119.
- [21] ZAMIR S W, ARORA A, KHAN S, et al. CycleISP: real image restoration via improved data synthesis[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2696-2705.
- [22] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition [C]//Proceedings of the 37th International Conference on Machine Learning, 2020: 3247-3258.
- [23] CHAI L, BAU D, LIM S N, et al. What makes fake images detectable? Understanding properties that generalize[C]//Proceedings of the 16th European Conference on Computer

- Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 103-120.
- [24] JU Y, JIA S, KE L, et al. Fusing global and local features for generalized ai-synthesized image detection[C]//Proceedings of the 2022 IEEE International Conference on Image Processing, 2022: 3465-3469.
- [25] ZHONG N, XU Y, QIAN Z, et al. Rich and poor texture contrast: a simple yet effective approach for ai-generated image detection[J]. arXiv:2311.12397, 2023.
- [26] CHIERCHIA G, POGGI G, SANSONE C, et al. A Bayesian-MRF approach for PRNU-based image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(4): 554-567.
- [27] SCHERHAG U, DEBIASI L, RATHGEB C, et al. Detection of face morphing attacks based on PRNU analysis[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2019, 1(4): 302-317.
- [28] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
- [29] TAN C, ZHAO Y, WEI S, et al. Learning on gradients: generalized artifacts representation for GAN-generated images detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12105-12114.
- [30] YU N, DAVIS L S, FRITZ M. Attributing fake images to GANs: learning and analyzing GAN fingerprints[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 2019: 7556-7566.
- [31] MARRA F, SALTORI C, BOATO G, et al. Incremental learning for the detection and classification of GAN-generated images[C]//Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security, 2019: 1-6.
- [32] JAVED K, SHAFAIT F. Revisiting distillation and incremental classifier learning[C]//Proceedings of the 14th Asian Conference on Computer Vision, Perth, Dec 2-6, 2018. Cham: Springer, 2018.
- [33] LOPEZ-PAZ D, RANZATO M. Gradient episodic memory for continual learning[C]//Advances in Neural Information Processing Systems 30, 2017: 6467-6476.
- [34] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. ICARL: incremental classifier and representation learning[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2001-2010.
- [35] WANG S Y, WANG O, ZHANG R, et al. CNN-generated images are surprisingly easy to spot for now[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8695-8704.
- [36] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
- [37] RICKER J, DAMM S, HOLZ T, et al. Towards the detection of diffusion model deepfakes[J]. arXiv:2210.14571, 2022.
- [38] GRAGNANIELLO D, COZZOLINO D, MARRA F, et al. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art[C]//Proceedings of the 2021 IEEE International Conference on Multimedia and Expo, 2021: 1-6.
- [39] 乔通, 陈彧星, 谢世闻, 等. 多色彩通道特征融合的 GAN 合成图像检测方法[J]. 电子学报, 2024, 52(3): 924-936.
- QIAO T, CHEN Y X, XIE S C, et al. GAN synthetic image detection using fused features in the multi-color channels [J]. Acta Electronica Sinica, 2024, 52(3): 924-936.
- [40] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[J]. arXiv:1710.10196, 2017.
- [41] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, 2017: 2223-2232.
- [42] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8789-8797.
- [43] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2337-2346.
- [44] WEST J, BERGSTROM C. Calling bullshit: the art of skepticism in a data-driven world[M]. New York: Random House, 2020.
- [45] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Advances in Neural Information Processing Systems 34, 2021: 8780-8794.
- [46] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[J]. arXiv:2112.10741, 2021.
- [47] HOLZ D. Midjourney[EB/OL]. [2024-05-13]. <https://www.midjourney.com/home/>.
- [48] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.
- [49] GU S, CHEN D, BAO J, et al. Vector quantized diffusion model for text-to-image synthesis[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10696-10706.
- [50] WUKONG. MindSpore[EB/OL]. [2024-05-13]. <https://xihe.mindspore.cn/modelzoo/wukong>.
- [51] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[J]. arXiv: 2204.06125, 2022.