基于扩散伪影对比学习的生成式图像检测方法

袁程胜[1,2*],陈金瑞[1],曹 燚[3],刘庆程[1],周志立[4],付章杰[1,2]

(1. 南京信息工程大学计算机学院、网络空间安全学院,江苏南京 210044;2. 南京信息工程大学数字取证教育部工程研究中心,江苏南京 210044;3. 无锡学院网络安全与信息化学院,江苏无锡 214105;4. 广州大学人工智能研究院,广东广州 510006)

摘 要: 随着以扩散模型为代表的生成式人工智能在视觉内容合成领域持续取得突破,其生成的图像在视觉真实感与内容多样性方面已逼近甚至部分超越真实摄影水平。然而,技术的快速发展也使生成式图像,特别是可能用于恶意目的的深度伪造内容的检测与鉴别任务变得日益复杂与严峻。现有大多数检测算法在受控的实验室环境下能够表现出较好的性能,但在开放的真实场景中,一旦面临训练数据与测试数据之间存在显著分布差异的情况,例如遇到未知的生成模型、未见过的图像风格或经过复杂后处理的伪造样本,这些方法的泛化能力与鲁棒性往往明显不足。为应对上述挑战,本文从困难样本分类的角度出发,提出一种基于扩散伪影对比学习(Contrastive Learning of Diffusion Artifacts,CLDA)的生成式图像检测方法,通过多模块协同优化,以提升模型对生成图像的检测精度与鲁棒性。首先,基于高质量扩散模型构造具有挑战性的生成样本,为模型训练提供更丰富的数据基础。随后,设计伪影增强模块,引入潜在空间跨域增强策略,通过基于余弦相似度加权的特征插值方法扩展伪造特征空间;同时结合域损失机制,引导编码器学习不同伪造域的鉴别性特征,避免模型对特定伪造模式过度依赖。进一步地,提出一种基于潜在空间边界的对比损失函数,通过动态权重聚焦于决策边界附近的困难样本对,以增强模型对真实图像、生成图像及反演图像间细微差异的辨识能力,并将该损失与二分类交叉熵损失相结合,构建统一的多目标优化函数。为验证本文所提方法的有效性,本文在GenImage与DRCT-2M两个公开数据集上进行了对比实验。实验结果表明,经过本文框架优化后的检测器,在GenImage数据集上的平均准确率提升1.1个百分点,在DRCT-2M数据集上的平均准确率提升4.8个百分点。此外,在图像缩放、JPEG压缩、高斯噪声等干扰场景下,本文方法仍保持较高的平均检测精度,其鲁棒性显著优于现有对比方法。

关键词: 生成式图像检测;扩散模型;伪造检测;图像取证;跨域增强;对比学习

基金项目: 国家自然科学基金(No.U22B2062,No.U23B2023,No.62102189)

中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0248-14

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250663

Generative Image Detection Based on Diffusion Artifact Contrast Learning

YUAN Chengsheng[1,2*], CHEN Jinrui[1], CAO Yi[3], LIU Qingcheng[1], ZHOU Zhili[4], FU Zhangjie[1,2]

(1. *School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu* 210044, *China*; 2. *Engineering Research Center of Digital Forensics Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, Jiangsu* 210044, *China*; 3. *School of Cybersecurity and Informationization, Wuxi University, Wuxi, Jiangsu* 214105, *China*; 4. *Institute of Artificial Intelligence, Guangzhou University, Guangzhou, Guangdong* 510006, *China*)

Abstract: With the continuous breakthroughs in generative artificial intelligence represented by diffusion models in the field of visual content synthesis, the generated images have approached or even partially surpassed real photographic levels in terms of visual realism and content diversity. However, the rapid development of this technology has also made the detection and identification of generated images—especially deepfake content that may be used for malicious purposes—increasingly complex and challenging. Most existing detection algorithms perform well in controlled laboratory environments, but in open real-world scenarios, once they encounter significant distributional differences between training and testing data—such as unknown generative models, unseen image styles, or forged samples subjected to complex post-processing—their generalization capability and robustness often exhibit notable deficiencies. To address these challenges, this paper proposes a generated image detection method based on contrastive learning of diffusion artifacts (CLDA) from the perspective of hard sample classification. The approach employs multi-module collaborative optimization to enhance the detection accuracy and robustness of the model for generated images. First, challenging generated samples are constructed using high-quality diffusion models to provide a richer data foundation for model training. Subsequently, an artifact enhancement module is designed, introducing a latent space cross-domain enhancement strategy. This strategy expands the forged feature space through feature interpolation weighted by cosine similarity, while incorporating a domain loss mechanism to guide

收稿日期:2025-07-29;录用日期:2025-12-26;责任编辑:覃怀银

*通信作者:袁程胜

the encoder in learning discriminative features across different forgery domains, thereby preventing the model from over-relying on specific forgery patterns. Furthermore, a contrastive loss function based on latent space boundaries is proposed, which employs dynamic weighting to focus on hard sample pairs near the decision boundary. This enhances the model's ability to discern subtle differences between real images, generated images, and inverted images. This loss is then combined with binary cross-entropy loss to construct a unified multi-objective optimization function. To validate the effectiveness of the proposed method, comparative experiments were conducted on two public datasets, GenImage and DRCT-2M. The experimental results demonstrate that the detector optimized by the proposed framework achieves an average accuracy improvement of 1.1 percentage points on the GenImage dataset and 4.8 percentage points on the DRCT-2M dataset. Additionally, under challenging scenarios such as image scaling, JPEG compression, and Gaussian noise, the proposed method maintains a high average detection accuracy, with its robustness significantly outperforming existing comparative methods.

Keywords: generated image detection; diffusion model; fake image detection; image forensics; cross-domain enhancement; contrastive learning

Foundation Item(s): National Natural Science Foundation of China (No.U22B2062, No.U23B2023, No.62102189)

0 引言

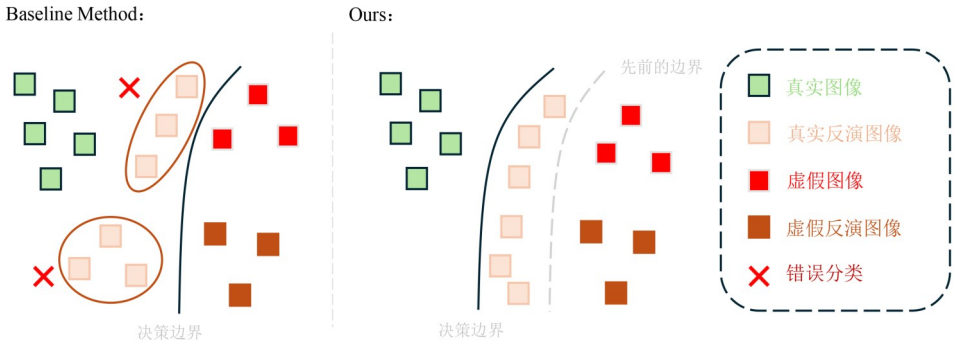
Deepfake技术作为人工智能领域的一项重要革新成果,因其能够生成极具真实感的伪造视频内容而受到广泛关注。该技术在娱乐、金融等众多领域展现出广阔的应用前景[1-2]。然而,其滥用亦可能带来伪造身份、虚假信息传播等一系列社会与伦理风险。因此,亟需深入审视并防范Deepfake技术可能引发的负面效应。近年来,生成模型(如生成对抗网络[3](Generative Adversarial Network,GAN),扩散模型[4])取得了显著进展,推动了图像合成技术从理论走向广泛应用。特别是去噪扩散模型在图像生成领域的突破,催生了一系列性能优异的新型生成架构。这些技术已被广泛应用于数字艺术、商业推广、媒体出版与娱乐产业等多个领域,成为内容创作的重要推动力。然而,其快速发展也引发了人们对生成内容所涉隐私与安全问题的深切关注[5]。值得警惕的是,这类技术存在被恶意利用的风险,可能被用于制造虚假信息、干扰政治进程、侵犯知识产权等非法活动。因此,开发高效、可控的生成图像检测技术,对构建安全可信的网络环境具有重大意义。当前,图像生成模型种类繁多且易于获取,这对检测方法的泛化能力提出了严峻挑战。理想的图像检测器不仅需准确识别来自已知模型的生成图像,还应具备对未见过的新模型生成图像的识别能力。尽管现有视觉取证技术在虚假媒体识别方面已取得显著成果,但大多仍局限于语义、物理层面或特定伪造场景下的统计差异分析。即使在扩散模型生成图像的检测任务中能够达到较高准确率,现有方法通常仅针对特定模型,依赖其独特的伪影特征[6]。一旦扩散模型的结构发生改变,这类检测方法的性能往往会显著下降。最近,Wang等人[7]提出了一种基于扩散重建误差的图像检测方法。该方法基于一个关键假设:相较于真实图像,扩散模型生成的图像更容易被同一模型准确重建。实验结果表明,该方法在面对不同结构的扩散模型时,表现出良好的泛化能力。当前,大多数生成模型的研究重点集中于生成图像复杂细节的刻画,而在一定程度上忽略了生成图像与真实图像在基础内容层面的一致性。为此,本文提出了一种基于扩散伪影对比学习(Contrastive Learning of Diffusion Artifacts,CLDA)的框架,该框架借鉴扩散重建误差的核心思想,通过扩散模型对真实图像进行重建生成反演样本。这类样本在视觉上与真实图像几乎难以区分,但仍保留了生成模型所遗留的细微伪影特征。利用此类具有挑战性的样本对现有检测方法进行训练,能够有效提升检测器的泛化能力,增强其对生成模型伪影的识别敏感度,使其在面对未知扩散模型生成的图像时仍具备良好的判别性能。此外,本文还设计了一种伪影增强模块,用于扩展伪造特征的空间分布。该模块能够引导模型构建更加鲁棒的决策边界,从而有效缓解因过度依赖特定伪造模式而导致的过拟合问题。具体流程如图1所示。为促使模型能够充分学习到真实图像、生成图像与反演图像之间的差异特征,本文基于潜在空间边界设计对比损失函数,在潜在空间中构建清晰的分类边界,从而进一步提升模型的检测性能。本文的主要贡献如下。

(1)本文提出了一种基于扩散伪影的对比学习框架,设计了一种伪影增强模块,通过引入潜在空间增强策略,对不同的潜在表征进行跨域语义增强,从而引导模型学习更具普适性的决策边界,并有效捕捉多类伪造特征之间共享的通用表征。此外,本文采用域损失策略,将每种伪造类型与真实类别均视为独立域处理,以引导编码器精准捕捉不同伪造域的判别性特征。

(2)本文设计了基于潜在空间边界的对比损失函数,利用反演图像中固有的指纹信息,计算不同样本对决策边界的损失贡献,引导检测器学习真实图像与

生成图像之间的细微差异。该方法能够有效提升现有检测器对扩散生成图像的识别精度与跨模型泛化能力。

(3)为验证本文方法在不同参数配置与真实场景下的性能表现,本文在GenImage和DRCT-2M两个数据集上进行了性能评估。结果表明,所提方法在泛化性与鲁棒性测试中均表现优异。此外,在多种数据分布和模型架构条件下同样取得了较好的性能。



注:本文方法能够让模型学习到更准确的决策边界。

图1 图像决策边界

Figure 1 Image decision boundary

1 相关工作

1.1 伪造图像生成

和变分自编码器(Variational AutoEncoder,VAE)自提出以后,便在计算机视觉领域扮演着先驱角色。然而,这两类模型在实现对生成图像内容的精准控制方面仍存在局限,这为新一代图像生成范式的发展提供了契机。以GAN为例,其中的无条件GAN以随机向量作为输入,无需依赖预定义的类别标签,因而能够生成高度开放且多样化的图像内容。GenImage数据集[8]是一个涵盖多种前沿生成模型合成图像的大规模基准,不仅覆盖广泛的图像类别,还专门设置了跨生成器泛化任务与退化图像检测任务,以模拟真实应用中的检测挑战。该数据集中的生成图像逼真度极高,达到了以假乱真的水平,充分体现了当前深度学习模型在捕捉与复现复杂图像分布方面的卓越能力。条件GAN则采用了与之不同的技术路径,该类模型无需依赖成对的标注数据,即可实现对图像生成过程的语义引导与精确控制。它能够有效学习不同图像域之间的风格转换规律与内容映射关系,在保持语义连贯性和结构合理性的基础上,实现图像特征的跨域迁移。在图像生成领域,扩散模型同样展现出巨大潜力,其借鉴了非平衡热力学和随机过程的理论框架,为深入理解与分析模型行为奠定了数学基础。其训练过程基于前向扩散与反向去噪的数学推导,直接采用均方误差等损失函数进行优化,有效规避了GAN因对抗训练所导致的不稳定问题。扩散模型具备良好的扩展性,能够灵活适配于多种数据类型与任务领域,如图像生成、文本生成与语音合成等[9]。其在医学图像重建、风格迁移、图像超分辨率等具体任务中,其性能表现尤为突出。Ho等人[10]提出的去噪扩散概率模型(Denoising Diffusion Probabilistic Model,DDPM)在生成质量与训练稳定性等方面均具有显著优势,已成为生成式图像发展历程中的重要里程碑。之后,Song等人[11]在DDPM的基础上引入非马尔可夫扩散过程,提出了去噪扩散隐式模型(Denoising Diffusion Implicit Models,DDIM),显著减少了生成高质量图像所需的采样步数。ADM[12]则通过改进模型架构,在无需依赖分类器指导的情况下,进一步提高了生成图像的视觉质量。在检测与理解方面,Tan等人[13]深入探讨了对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)的检测机制,通过将图像特征解码为文本并进行词频分析,揭示了其辨别生成图像的内在原理。Zhao等人[14]提出了一种统一的多模态控制框架,能够在单一模型中灵活融合边缘图、深度图、分割掩码等局部控制信号,以及CLIP图像嵌入等全局信息,实现了生成过程的高效组合与控制。Rombach等人[15]提出的潜在扩散模型(Latent Diffusion Model,LDM)通过引入交叉注意力机制和潜在空间表示,将扩散过程压缩至低维空间进行,不仅降低了计算成本,也使生成过程更集中于语义信息的引导。与GAN中生成器与判别器之间难以稳定的对抗训练相比,扩散模型在训练过程中表现出更强的稳定性和收敛性,有效避免了模式崩溃等常见问题。

1.2 伪造图像检测

近年来,生成图像检测的研究主要围绕GAN所

生成的图像,并形成了多种针对性方法。早期研究多依赖于人工设计的特征实现判别,如采用混合伪影等手工特征作为检测依据[16]。Vasilcoiu 等人[17]提出一种基于潜在轨迹嵌入的方法,通过建模潜在嵌入在多个去噪时间步中的演化轨迹,以捕捉细微且具判别性的伪造模式。Wang 等人[18]在一个大规模真实图像与合成图像数据集上系统采用数据增强策略,研究发现,使用经过 JPEG 压缩与模糊处理的 ProGAN 图像训练得到的简单卷积神经网络分类器,能够有效迁移至其他 GAN 生成图像的检测任务,显著提升了检测器对未知生成图像的泛化能力。Frank 等人[19]从频域视角分析指出,GAN 生成图像通常存在较明显的频域伪影,这主要源于其生成架构中上采样操作的固有特性。Yu 等人[20]则通过提取 GAN 模型的独特“指纹”特征,将其作为图像来源鉴别的重要依据。

然而,随着扩散模型的快速发展,该领域正面临一项重要挑战:目前仍缺乏泛化能力强、适用范围广的检测器,能够准确识别各类扩散模型生成的图像。在相关研究中,Radford 等人[21]利用自然语言监督信号,成功训练出性能优异的视觉模型。该研究在预训练阶段采用对比语言-图像预训练方法,并基于文本提示实现零样本迁移学习。Ojha 等人[22]则使用预训练的 CLIP 模型作为特征提取器,结合最近邻分类器,显著提升了对不同图像生成方法的泛化能力。与此同时,Li 等人[23]从模型特定噪声印记的角度出发,提出一种噪声印记模拟器,旨在捕捉不同生成模型输出图像中的固有噪声模式。该方法融合了基于噪声印记提取器衍生的噪声特征,以及其他与生成图像检测相关的视觉特征,设计了判别能力更强的检测器。Liu 等人[24]通过频域分析发现,真实图像的噪声模式呈现高度一致性,而生成图像则存在显著差异。基于此,他们提出通过检验图像是否符合真实图像的噪声分布规律,实现图像的真伪鉴别。此外,Guarnera 和 Guo 等人[25-26] 分别提出了不同的多层级分析方法。这些方法不仅能够有效区分由 GAN 与扩散模型生成的图像,还具备检测局部篡改伪造图像的能力。Qian 等人[27]提出了一种新颖的频率人脸伪造网络(Face forgery detection by mining frequency-aware clues, F3Net),该网络利用频率感知的伪造线索挖掘细微的伪造模式,从局部频率统计信息中提取高级语义特征,从而精准刻画真实面孔与伪造面孔在频率统计特性上的差异。Liu 等人[28]对底层纹理统计分析后,提出了 Gram-Net 架构,通过捕捉全局图像纹理特征,构建了一种对生成图像具有强鲁棒性的检测方法。Tan 等人[29]指出,在 GAN 或扩散模型生成的合成图像中,上采样算子会显著增强图像像素间的局部相互依赖性。基于这一发现,他们引入了“相邻像素关系”的概念,以系统性捕捉上采样操作引起的通用结构伪影。Luo 等人[30]提出了一种基于潜在重构误差引导的特征细化方法(Latent reconstruction error based method for diffusion-generated image detection, LaRE2),该方法从空间和通道两个维度对生成图像的检测特征进行优化,显著提升了扩散生成图像的检测性能。

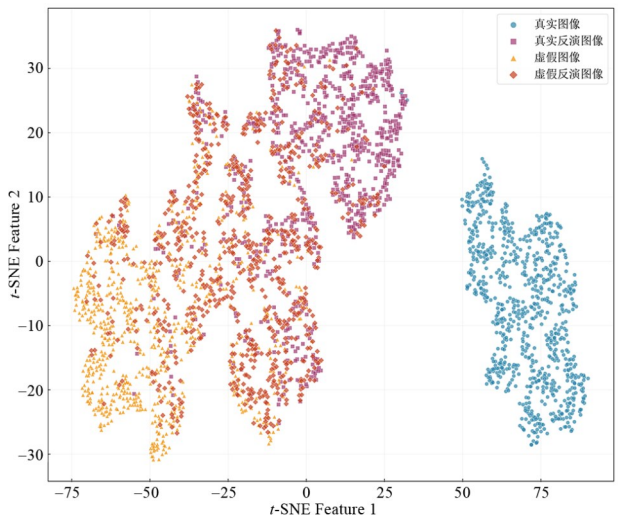
现有研究通过挖掘频域伪影、噪声印记以及上采样引发的局部像素关系等多维特征,并引入基于重建误差的检测视角,为生成图像识别奠定了重要基础。然而,多数方法过度依赖特定生成模型的显式指纹,导致其在面对未知扩散模型时识别性能显著下降,泛化能力受限;同时,基于多步迭代重建的方法计算开销巨大,难以兼顾检测效率与鲁棒性。针对上述挑战,本文提出一种面向生成图像的检测框架,利用高效的反演技术构建困难样本对,并通过伪影增强模块与潜在空间边界对比损失,引导模型学习更具普适性的决策边界,从而显著提升对未知生成模型的识别能力。

2　主要架构

本文提出的扩散伪影对比训练框架采用双阶段流程,包括图像重建反演与模型训练两个核心步骤。在图像重建阶段,利用预训练的扩散模型分别对真实图像与生成图像进行重构,构建多样化的图像样本库。该过程充分发挥扩散模型的强大生成能力,通过对原始图像进行二次重构,生成包含“真实图像反演结果”与“生成图像反演结果”的混合样本集合。在模型训练阶段,采用四类图像样本参与训练:原始真实图像、基于真实图像反演得到的图像、扩散模型生成的伪造图像,以及基于伪造图像反演得到的图像。本文设计了伪影增强模块,通过潜在空间增强扩展伪造特征的表示空间,并联合分类损失与伪造域损失,引导编码器有效学习不同伪造域的特征表示。此外,训练过程中利用图像重建阶段生成的多类别样本,计算基于潜在空间边界的对比损失。通过对比反演样本与真实样本在潜在空间中的边界分布差异,模型能够学习到对未知伪造手段更具鲁棒性的本质差异特征。最终,通过多种损失函数的协同优化,实现将真实图像准确归类为真实类别,并将各类伪造图像判定为虚假类别。如图 2 所示,本文模型具有良好的判别能力。

2.1　图像重建

基于图像重建反演的数据增强方法在现有研究中已被广泛采用,然而现有方法如 DIRE[7]在特征提取时需经过多步 DDIM 采样,导致图像重建过程耗时较长、效率受限。值得注意的是,真实图像的重建难



注:本文方法能够让模型具有更好的判别能力。

图2 t-SNE簇中心嵌入的可视化

Figure 2 Visualization of t-SNE cluster center embeddings

度通常高于生成图像,这进一步制约了该类方法的应用效率。为此,本文在图像重建阶段采用了SDv1.4模型,通过调控潜在变量以增强扩散过程的表达能力,可高效生成多样化图像样本。与DIRE相比,本文方法在保持重建质量的同时,显著降低了计算开销,缩短了图像重建时间。具体而言,本阶段旨在利用扩散模型的反演能力,对原始图像进行重构与再生成,从而构建一个包含真实图像反演结果与生成图像反演结果的混合样本集,为后续任务分类器训练提供数据基础。重建过程以前向扩散为起始,逐步向图像中添加噪声;随后通过反向去噪过程,迭代降低噪声水平,逐步恢复出结构合理的反演图像。该机制有效保证了生成样本的多样性与重建质量。

在人工智能生成图像检测领域,基于扩散模型的框架凭借其独特的噪声添加与去噪机制,已成为提取图像深层表征的有效途径。本文将图像编码至潜在空间以进行后续分析,其核心流程围绕正向扩散与反向重建两个关键环节展开。

具体而言,首先将输入图像x通过预训练的编码器映射至潜在空间,得到对应的向量表示x_0。随后,在正向扩散过程中,依据预设的噪声调度策略,逐步向x_0中添加噪声,生成不同阶段的带噪数据。正向扩散过程如下:

$$x_t=\sqrt{\bar{a}_t}\,x_0+\sqrt{1-\bar{a}_t}\,\epsilon \tag{1}$$

其中,当变量$t=0,1,\cdots,T$时,随机变量ϵ服从正态分布,即$\epsilon\sim N(0,1)$。在上述公式中,x_0表示初始的数据;x_t为经过t步扩散处理后得到的含噪声数据;a_t为预先设定的噪声时序参数,满足$\bar{a}_t=\prod_{s=1}^{t}a_s$。

正向扩散过程作为模型的起始阶段,通过对原始潜在表示施加高斯噪声,使其逐渐接近纯噪声状态。如图3所示,相较于DIRE方法中依赖多步迭代的采样流程,本文提出的方法基于马尔可夫过程与高斯分布的特性,能够直接从初始数据x_0推导得出第t步扩散后的数据x_t,降低了计算时间开销,计算公式如下:

$$q(x_t|x_0)=N\left(x_t;\ \sqrt{a_t}\,x_0,(1-a_t)I\right) \tag{2}$$

在完成正向扩散过程与模型训练后,反向重演便成为图像重建的关键步骤。该过程基于训练好的去噪神经网络,沿正向扩散的逆序逐步从带噪数据中恢复原始图像的特征信息。

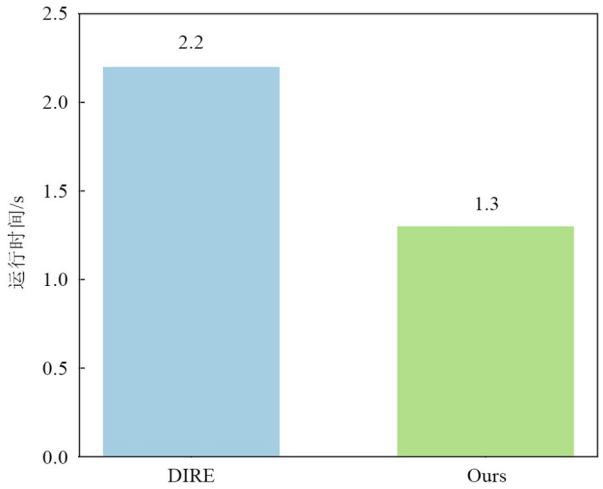


图3 利用扩散模型进行图像重建反演时间成本比较

Figure 3 The time cost of image reconstruction inversion is compared using the diffusion model

本质上,它是正向扩散的逆过程,其目标是将正向过程中逐步叠加噪声所形成的近似纯噪声数据,重构为清晰的原始图像。与正向扩散的随机性不同,反向重演是一个确定性的计算流程,其核心在于逐步去除噪声,恢复出原始图像的潜在结构,这一过程可概括表述为

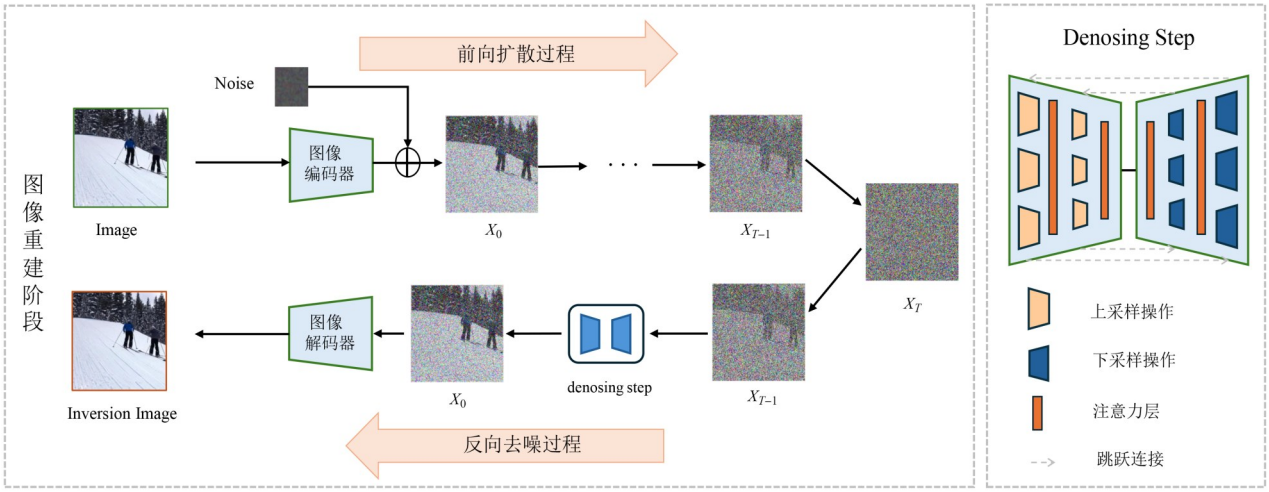
$$x_{t-1}=\frac{1}{\sqrt{a_t}}\left(x_t-\frac{1-a_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta\left(x_t,t\right)\right)+\sigma_t\epsilon \tag{3}$$

在反向重建过程中,本文将方差设定为特定值$\sigma_t^2=1-a_t$。其中,$\epsilon_\theta(x_t,t)$是由参数θ参数化的去噪神经网络所预测的噪声,而$\epsilon\sim N(0,1)$为高斯噪声。通过迭代执行上述公式,模型能够从完全噪声化的状态逐步还原至初始的潜在表示。在具体实现中,利用SDv1.4模型对带噪数据进行去噪处理,在反向重演过程中不断修正潜在表示,最终得到高度噪声化的潜在表示x_0'。为获得最终的重构图像,需要通过解码器将潜在空间中的表示映射回原始图像空间。扩散重建

模型的训练目标是通过神经网络学习去噪能力,以实现从噪声图像中恢复原始图像的过程。该过程的优化目标如式(4)所示:

$$L(\theta)=\left\|\epsilon-\epsilon_\theta(x_t,t)\right\|^2=\left\|\epsilon-\epsilon_\theta(\sqrt{\bar{a}_t}\,x_0+\sqrt{1-\bar{a}_t}\,\epsilon,t)\right\|^2 \tag{4}$$

在训练过程中,扩散步骤 t 和噪声 ϵ 均通过随机抽样的方式获取。模型通过最小化预测噪声 $\epsilon_\theta(x_t,t)$ 与真实噪声 ϵ 之间的均方误差,不断优化参数 θ,从而提升去噪神经网络的性能。上述操作不仅能够高效提取图像的深层特征,也为后续的生成图像检测任务提供了关键的数据。如图4所示,本文在稳定扩散框架下进行图像重建,该过程基于一个条件引导的扩散模型。



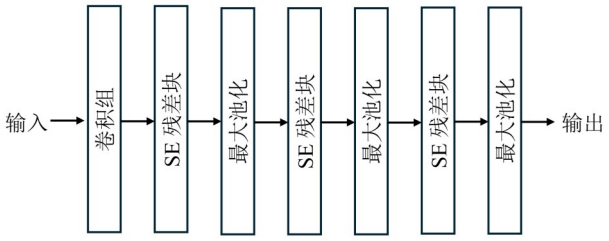
注:利用SDv1.4对图像进行重建反演,生成新的数据集样本。

图4　利用扩散模型进行图像重建反演

Figure 4　Image reconstruction inversion is carried out using the diffusion model

2.2　对比学习过程

随着扩散模型在图像生成领域日益普及,其生成图像的视觉逼真度持续提升,对传统生成式图像检测方法构成了显著挑战。本文旨在设计与训练高效检测器,以实现对扩散模型生成图像的精准判别。研究发现,扩散模型通过其特有的噪声添加与去噪机制进行图像重构,所生成图像具有更为复杂的特征表达,从而进一步加大了检测难度。为此,本文提出一种基于扩散伪影对比学习的检测框架,如图5所示。在图像生成部分,将输入图像记为 $x\in\mathbf{R}^{W\times H\times C}$。随后,该图像被送入主干网络以进行特征提取。主干网络采用经典的层级式架构,通过堆叠多层卷积与池化操作,逐步压缩图像的尺寸,同时增强其特征表达能力。接着,将经骨干网络处理后的特征 F 输入到预测层进行训练。预测层模块由三个残差块、一个卷积层和四个最大池化层组成。如图5所示,残差块能有效提取图像的固有特征,增强特征表达的判别力。在参数方面,第一个卷积组设置32个卷积核,三个残差块的卷积核数量依次为32、64和128。随着网络层级的加深,卷积核数量不断增加,使网络能够学习到从低级到高级的层次化特征表示。具体而言,浅层网络主要捕捉边缘、纹理和颜色等基础视觉信息,随着层级推进,网络逐渐提取出更具语义意义的形状结构与抽象特征,从而为图像检测任务提供多层次、高判别力的特征表示。



注:残差块能够提取图像固有的特征信息,有效提升特征的有效性。

图5　残差块设计框架

Figure 5　Residual block design framework

现有伪造检测模型普遍存在缺陷,易对特定伪造域的显性特征(如GAN生成的纹理重复、扩散模型的局部模糊)产生过拟合,面对新型伪造手段检测精度不高,泛化性严重不足。基于图像重建阶段生成的反演样本,设计了伪影增强模块,通过潜在空间跨域增强与域特征强制分离的设计,扩大伪造样本的潜在空间分布范围,使模型聚焦于反演样本与真实样本的边界差异,而非局限于特定域特征。具体而言,将伪造域集合设置为 $D=\{D_1,D_2,D_3\}$,对每个伪造域 D_i 的潜

在表征 $F_i \in \mathbf{R}^{W\times H\times C}$，随机选取另一个不同伪造域 D_k 的潜在表征 F_k，构建跨域样本对。通过引入域特征相似度感知的权重因子 α，计算 F_i 与 F_k 的特征相似度：

$$\alpha = \mathrm{clip}\left(\rho(F_i, F_k) + \epsilon, 0.1, 0.9\right) \tag{5}$$

其中，$\rho(F_i, F_k)$ 为余弦相似度；ϵ 为随机扰动，旨在增强混合结果的多样性。为扩大伪造特征空间并防止模型对特定伪造特征产生过拟合，引入了潜在空间增强技术对不同的潜在表征 F 进行跨域增强：

$$\tilde{F}_i = aF_i + (1-a)F_k, \quad i \neq k \in \{1,2,3\} \tag{6}$$

其中，i 和 k 为不同的伪造域标识；a 表示 $[0,1]$ 范围内随机采样的权重系数；$\tilde{F}_i$ 表示跨域增广样本。该方法在多个伪造域之间进行特征层面的融合，构建了一种动态混合策略。其通过计算特征相似度以自适应地调节混合权重，主动扩大了伪造样本在潜在空间中的分布范围，迫使模型不能只记忆单一域的特征，使生成的增强样本既能够扩大伪造空间，又保持伪影特征的物理合理性与语义一致性。

本文还针对不同伪造域设计了域损失策略，引导编码器有效学习不同伪造域的判别性特征，将每种伪造类型和真实类别视为独立域，促使模型聚焦于类别间的“临界差异”，例如生成图像在边缘区域存在的纹理异常或结构缺陷，从而增强对真实图像与生成图像细微差异的识别能力。具体实现中，将输入图像 x_i 编码至潜在空间得到其特征表示 r_i 中，随后通过一个多类别分类器估计该特征属于各域的置信度分数 $s_i \in \mathbf{R}^{B\times(m+1)}$。利用 softmax 函数将置信分数 s_i 转化为似然值 v_i，即 $v_i = \mathrm{softmax}(s_i)$，然后计算域损失如下：

$$L_{\mathrm{domain}} = -\frac{1}{B\times(m+1)} \times \left[\log\left(1-(v_0)_j\right) + \sum_{i=1}^{m}\log\left((v_i)_j\right)\right] \tag{7}$$

其中，$(v_i)_j \in \mathbf{R}$ 表示第 j 个特征被分类为域 i（0 表示真实）的伪造概率。域损失函数将真实与不同伪造类型视为不同的域，并设计损失函数强制编码器学习区分这些域。其目的不是最终的多分类，而是作为一种特征提炼手段，引导编码器捕捉到各类伪造域与真实域之间最关键的“临界差异”。通过上述设计，伪影增强模块能够促使模型聚焦于关键伪影特征，为后续损失函数的优化提供更具区分度的潜在表征。

为进一步增强模型对真实图像、生成图像及其反演结果间差异的辨别能力，本文设计了一种基于潜在空间边界的对比损失函数。具体而言，对于正样本对（即相似样本），该损失函数通过减小其在特征空间中的距离，促使同类样本聚合；对于负样本对（即不匹配样本），则通过增大彼此距离，推动异类样本分离，从而在潜在空间中形成更具判别性的分类边界。为强化边界区域样本对训练过程的影响，本文通过引入边界权重因子 ω_n 来聚焦边界区域样本对，自适应地为位于边界区域的“难区分样本对”分配更高的权重。这使得模型训练的重心集中在那些最富信息量、最能体现真伪细微差别的样本上。该因子作为第 n 个样本对的动态权重，用于量化该样本对在潜在空间边界中的贡献度，定义为

$$\omega_n = \exp\left(-\frac{d_n^2}{2\sigma^2}\right) \tag{8}$$

其中，σ 是尺度参数，用于控制边界区域的宽窄；d_n 是样本对 (a_n, b_n) 到潜在空间决策边界的距离，定义为

$$d_n = \left|\mathrm{Dist}(a_n, C_1) - \mathrm{Dist}(b_n, C_2)\right| \tag{9}$$

其中，C_1 和 C_2 分别记为两类样本的潜在空间中的聚类中心；$\mathrm{Dist}(x, C)$ 表示样本 x 到聚类中心 C 的欧氏距离。当样本对位于决策边界附近时，损失贡献被最大化；当样本对远离决策边界时，损失贡献被抑制。该设计不仅简化了损失计算复杂度，同时增强了模型对类别间细微差异的感知能力，其数学公式表达为

$$\begin{aligned} L_{\mathrm{contrastive}} = \frac{1}{2N}\sum_{n=1}^{N}\omega_n\Big[y_n\left\|a_n - b_n\right\|^2 + \cdots + (1-y_n) \\ \cdot\max\left(\mathrm{margin} - \left\|a_n - b_n\right\|, 0\right)^2\Big] \end{aligned} \tag{10}$$

其中，N 是样本对总数。对于每一对样本，设有一个二值标签 y_n，用于标识该样本对是否匹配（1 表示匹配，0 表示不匹配）。a_n 和 b_n 分别记为样本对的特征表示，$\left\|a_n - b_n\right\|$ 为两者特征间的欧氏距离。margin 是一个预设的阈值，默认值是 1.0，用于约束负样本对之间的最小间隔。通过精细化调整 margin 参数，可灵活控制负样本对在潜在空间中的分离程度，该机制不仅增强了模型对复杂数据分布的适应能力，也显著提升了分类性能与泛化性。模型将重点学习决策边界样本对的差异，尤其是生成图像与真实图像在边界区域所呈现的细微纹理差异。该损失函数通过强化对边界区域样本的关注，有效缓解了高度不平衡分割任务中区域损失计算所面临的数值不稳定及梯度主导问题。

交叉熵损失函数是深度学习中广泛用于二分类任务的损失度量方法，其通过计算模型预测的概率分布与真实标签分布之间的 KL 散度（Kullback-Leibler Divergence，又称相对熵）差异，为模型参数优化提供方向。在训练过程中，通过梯度下降算法最小化该损失，模型能够逐步学习从输入特征到输出类别的非线性映射关系，从而在测试阶段实现更高的分类准确率。交叉熵损失计算如下：

$$L_{\text{cross-entropy}}=-\sum_{n=1}^{N}\left[y_n\cdot\log(p_n)+\cdots+(1-y_n)\log(1-p_n)\right] \tag{11}$$

其中,N表示样本总数;y_n为第n个样本的二值标签(1表示正类,0表示负类);p_n是模型预测第n个样本为正类的概率。为综合考虑对比学习、伪影增强模块和分类任务的需要,模型的总体优化目标是对比损失、域损失和二分类交叉熵损失和加权组合,具体如下:

$$L_{\text{total}}=\lambda_1 L_{\text{contrastive}}+\lambda_2 L_{\text{domain}}+\cdots+(1-\lambda_1-\lambda_2)L_{\text{cross-entropy}} \tag{12}$$

其中,参数λ取值于[0,1)区间,用于动态调节不同损失项之间的平衡,在本文实验中,λ_1的默认值为0.3,λ_2的默认值为0.3。本文所提模型的参数量约为93.3 M,与现有主流方法相比,UnivFD[22]采用CLIP-ViT-L/14作为骨干网络(参数量约420 M),本文模型通过高效的网络结构设计,降低了计算开销与部署成本,尽管其参数量高于DIRE[7](约88 M),但在检测精度上实现了性能超越,展现出更优的权衡特性。如图6所示,通过反演数据增强、跨域伪影混合和边界对比学习三重机制,模型不再依赖记忆特定伪造特征,而是学会了捕捉由生成模型底层机理所导致的、更具普适性的伪影模式,对未知模型生成的图像具有更强的识别能力。虽然流程是双阶段的,但第一阶段的高效反演为第二阶段提供了高质量、高多样性的训练数据,从整体上提升了模型的训练效率和最终性能,避免了在单一模型上过拟合。总而言之,本文所提方法的核心优势在于它从一个全新的生成过程痕迹的视角出发,通过一套精心设计的主动式框架,最终训练出一个能抓住本质、举一反三的深度检测模型。

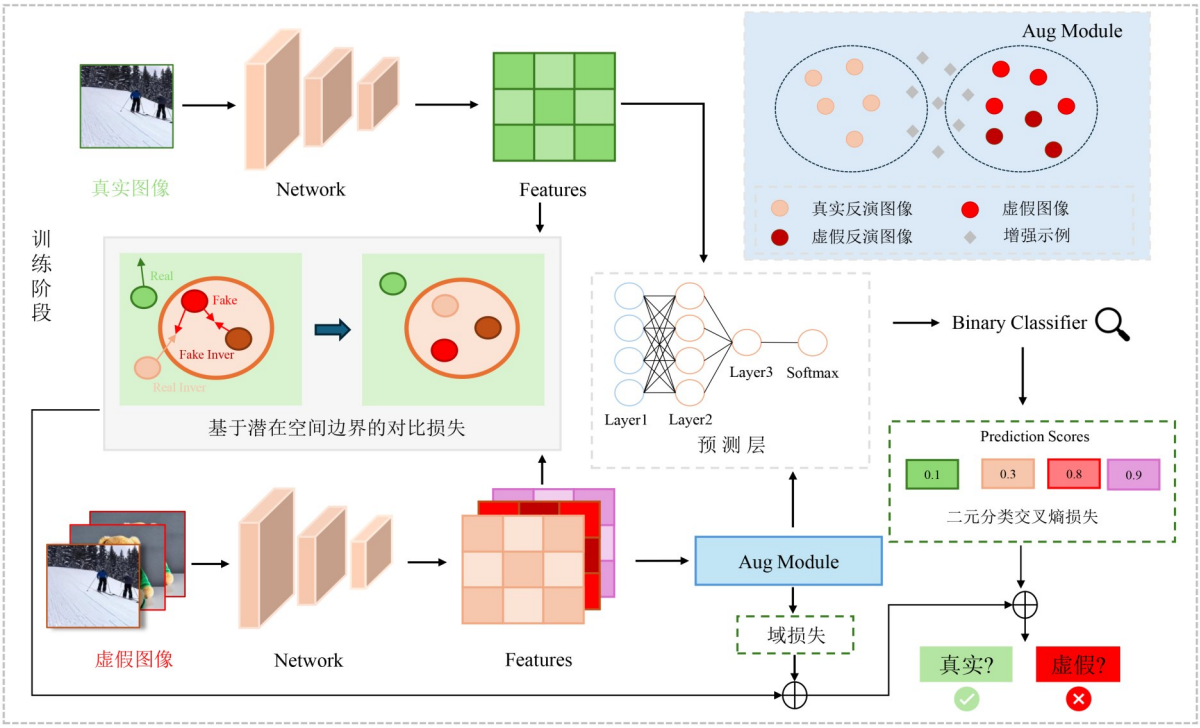


图6　基于扩散伪影的对比学习框架

Figure 6　A contrastive learning framework based on diffusion artifacts

3　实验

3.1　实验设置

在模型架构的设计阶段,本文选择ConvNeXt-Base模型作为网络的核心主干架构。训练过程中,为了有效增强数据的多样性,输入网络的图像会先进行随机裁剪,调整至224×224像素的尺寸,同时以50%的概率对图像进行水平翻转操作,以增强数据多样性。此外,为了生成与真实图像高度相似、难以区分真伪的反演图像,还引入了Stable Diffusion反演重建技术,并将默认的扩散步数设置为S=20。在测试阶段,图像统一中心裁剪为224×224像素的尺寸,以此确保测试过程的规范性和结果的准确性。在训练检测模型时,为了更好地捕捉全局特征,本文采用了Adam优化器,学习率设定为2×10^{-4},批大小则设置为32。本研究的实验基于Nvidia GeForce RTX 3090 GPU开展,并借助PyTorch框架进行各项测试工作。在评估检测器性能方面,将准确率(ACCuracy,ACC)

作为主要评估指标，同时设定阈值为0.5来计算ACC。这一设定能够直观反映模型的整体预测能力，从而将所提出的方法与先前的研究成果进行公平且有效的对比。

本文选用GenImage数据集[8]和DRCT-2M[9]这两个具有代表性的大规模生成图像检测数据集进行实验。GenImage数据集规模庞大，包含超过268万张图像，其中真实图像约133万张，均取自ImageNet数据集，而AI生成的虚假图像多达135万张，涵盖八个主流生成模型（包括七个扩散模型与一个GAN模型）。DRCT-2M数据集包含文本到图像和图像到图像两类生成内容，包含16种生成式模型，每种图像约12万张，其输入提示来源于MSCOCO数据集，并借助ControlNet模型完成生成。这些图像高度逼真，风格多样，给检测任务带来了较大的挑战。在本实验中，本文使用GenImage中的SDv1.4子集来训练模型。此外，还利用SDv1.4模型生成反演图像，以增强数据集的多样性和复杂性。这两个数据集包含了多种生成方式和后处理技术，能够全面评估模型在不同情境下的性能。

为更清晰展示本文所提方法的性能，本文还在结果表格中将最优平均准确率以加粗形式标出，并与当前多种先进方法进行系统比较，这些方法包括DIRE[7]（CVPR2023）、DRCT[9]（ICML2024）、C2P-CLIP[13]（AAAI2025）、LATTE[17]（arXiv2025）、CNNSPOT[18]（CVPR2020）、ResNet50[21]（ICML2021）、UnivFD[22]（CVPR2023）、F3Net[27]（ECCV2020）、GramNet[28]（CVPR2020）、NPR[29]（CVPR2024）、LarE2[30]（CVPR2024）、Conv-B[31]（CVPR2022）等。

3.2 泛化性能评估

在图像生成检测领域，泛化能力是评估模型性能的关键指标之一，直接影响模型在面对未知生成样本时的识别准确率，进而决定检测系统的实际应用价值。近年来，扩散模型所生成图像的质量与多样性不断提升，在此背景下，如何准确、高效地评估检测方法的泛化性能，已成为学术界与工业界共同关注的关键问题。为系统且全面地验证本文所提出方法在扩散模型生成图像检测任务中的准确性与泛化能力，本文选用GenImage和DRCT-2M两个具有代表性的大规模数据集进行实验。这两个数据集涵盖了多种生成技术和后处理方法，能够全面评估模型在不同场景下的表现。

近期，新发布的GenImage数据集被设置成8个子集，所有图像均由当前主流生成器合成。在训练阶段，本文采用该数据集中的SDv1.4子集对所提模型进行训练。为构建生成图像与真实图像的对比训练样本，同时引入MSCOCO数据集作为真实图像来源，并利用SDv1.4模型对其实施反演处理，以增强训练数据的多样性与挑战性。在测试阶段，本文利用GenImage数据集其余的7个子集进行测试。这些子集所包含的图像，既涵盖了由多种生成器生成的图像，也有经过不同后处理方式处理的图像，能够对模型在跨数据集以及跨生成技术场景下的泛化能力进行全面评估。通过在这些风格各异的子集上进行测试，能够更精准地评估模型在面对未知数据时的表现。

实验结果如表1显示，本文方法在GenImage数据集上表现优异，平均准确率达83.6%，较当前最优方法进一步提升1.1个百分点。在主流扩散模型检测任务中，CLDA在Midjourney数据集上取得94.3%的准确率，略低于DRCT的94.6%，但仍表现出较强的检测能力。在SDv1.4和SDv1.5数据集上，本文方法准确率接近100%，显示出对高保真生成模型优异的识别性能。对比方法方面，UnivFD与LarE2虽在部分模型上表现良好，但在Midjourney、SDv1.4等主流生成模型上的准确率明显低于本文方法。这一结果表明，CLDA框架对扩散模型生成图像具有较好的通用性，也体现了其出色的检测性能。在面对多样化生成技术时，CLDA仍能保持稳定的高准确率，展现出良好的泛化能力。为比较不同扩散模型对生成反演样本检测效果的影响，本文还额外采用SDv2.0作为图像重建模型进行结果实验，其实验设置与SDv1.4一致，仅替换重建模型。如表2所示，无论基于何种扩散先验，本文所提出的检测器均表现优异。其中，基于SDv1.4的检测效果最佳，这主要由于GenImage数据集中包含大量源自SDv1系列模型的生成样本，使模型在该系列上具备更强的特征适配性。

为全面评估本文方法对不同扩散模型生成图像的泛化能力，本文在新发布的DRCT-2M数据集上进行了系统测试，重点验证其对高质量生成图像的识别鲁棒性。实验结果如表3所示，本文方法在检测新型生成模型生成的图像时表现出卓越的泛化性能，平均准确率达到88.3%，优于所有现有先进方法，较之前最优结果提升4.8个百分点，充分证明了本方法的有效性与泛化优势。对于主流扩散模型，本文所提方法在LDM、SD v1.4/1.5上取得了99.9%的准确率，在SD-Turbo上也达到86.2%，显示出对成熟扩散模型生成图像的高判别能力。对比方法中，UnivFD在传统模型中，例如SD v1.4上表现尚可，但在新型模型，如LCM-SDXL、SDv2-DR上的准确率明显低于本文方法，反映出本文方法提出的扩散伪影对比学习能够更好地捕捉这些新模型的共有伪影特征，进而提升了任务

表1　在GenImage数据集上,不同方法对应的性能测试　　单位:%

Table 1　Performance tests corresponding to different methods on the GenImage dataset　　unit: %

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	AVG
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Ours	94.3	99.6	99.6	66.4	73.3	99.5	76.8	59.4	**83.6**
CNNSPOT[18]	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
F3NET[27]	50.0	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
ResNet50[21]	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
GramNet[28]	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
UnivFD[22]	91.5	96.4	96.1	58.1	73.4	94.5	67.8	57.7	79.4
DIRE[7]	60.2	99.9	99.8	50.9	55.0	99.2	50.1	50.2	70.7
NPR[29]	66.4	98.9	98.8	55.6	67.7	97.6	60.3	54.2	74.9
LarE2[30]	66.4	87.3	87.1	66.7	81.3	85.5	84.4	74.0	79.1
DRCT[9]	94.6	99.8	99.8	61.8	65.9	99.9	74.8	58.8	81.9
LATTE[17]	71.3	79.3	81.8	82.8	92.8	82.0	82.9	87.8	82.5
C2P-CLIP[13]	56.6	77.5	76.9	71.6	73.5	84.4	73.7	85.9	75.0

注:加粗为对比最优结果。

表2　在GenImage数据集上,利用不同扩散模型重建反演样本对比实验效果　　单位:%

Table 2　On the GenImage dataset, the inversion samples were reconstructed using different diffusion models to compare the experimental effects　　unit: %

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	AVG
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Ours	94.3	99.6	99.6	66.4	73.3	99.5	76.8	59.4	**83.6**
Ours(SDv2)	98.2	97.8	97.8	60.0	60.0	95.9	61.3	52.3	77.9

注:加粗为对比最优结果。

表3　在DRCT-2 M数据集上,不同方法对应的性能测试　　单位:%

Table 3　Performance tests corresponding to different methods on the DRCT-2 M dataset　　unit: %

Method	LDM	SD v1.4	SD v1.5	SD v2	SD XL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	AVG
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Ours	99.9	99.9	99.9	94.9	73.9	74.7	86.2	62.5	99.5	66.5	99.9	89.3	73.0	100	99.4	92.8	**88.3**
CNNSPOT[18]	99.9	99.9	99.9	97.5	66.3	86.6	86.1	72.4	98.2	61.7	97.9	85.8	82.8	60.9	51.4	50.3	81.1
F3NET[27]	99.8	99.8	99.8	88.7	55.8	87.4	68.3	63.7	97.4	54.9	98.0	72.4	82.0	65.4	50.3	50.3	77.1
ResNet50[21]	99.0	99.9	99.9	94.6	62.1	91.4	83.4	64.4	98.9	57.4	99.7	80.7	82.1	65.8	50.7	50.5	80.0
GramNet[28]	99.4	99.0	98.8	95.3	62.6	80.7	71.2	69.3	93.1	57.0	90.0	75.6	82.7	51.2	50.0	50.1	76.6
UnivFD[22]	98.3	96.2	96.3	93.8	91.0	93.9	86.4	85.9	90.4	89.0	90.4	81.1	89.1	52.0	51.0	50.5	83.5
C2P-CLIP[13]	97.2	96.3	93.3	84.3	53.5	67.9	74.7	60.6	90.1	66.1	86.9	68.7	77.8	67.2	57.1	56.7	74.9
Conv-B[31]	99.9	99.9	99.9	95.8	64.4	82.0	80.8	60.8	99.3	62.3	99.8	83.4	73.3	61.7	51.8	50.4	79.1
NPR[29]	98.6	99.8	98.6	94.7	67.7	83.3	82.1	63.4	98.3	63.3	98.2	81.1	73.0	68.8	64.4	72.1	81.7

注:加粗为对比最优结果。

的检测性能。同时,所设计的伪影增强模块可引导编码器聚焦于图像边界区域的纹理缺陷,促进模型学习不同生成模型的共性伪影特征,避免对特定伪造特征的过拟合,进一步增强模型的泛化能力。

3.3 鲁棒性能评估

此外,本节将对所提方法与当前主流的几个检测方法在应对模糊处理与图像压缩后的鲁棒性进行了对比分析。实验在GenImage数据集上进行,测试图像(包括真实图像与生成图像两类)分别经过尺寸缩放(缩放比例为0.6、0.8、1.0、1.2、1.4)、JPEG压缩(质量因子选取60、70、80、90、100)以及高斯噪声(标准差为0.5、1.0、1.5、2.0)处理。实验选取Conv-B、UnivFD、NPR和CLDA四种检测器进行鲁棒性评估。

所有方法均在统一的框架下完成训练,并采用相同的数据增强策略。如图7所示,本文所提方法在鲁棒性方面表现优异。具体来说,在图像尺寸缩放操作过程中,该方法受到的干扰极小,平均检测准确率始终稳定保持在97.7%的较高水平。在对图像进行JPEG压缩处理后,本文所提方法的平均检测准确率进一步提升至98.5%,在所有方法中位居首位。相比之下,未针对鲁棒性进行专门设计的方法,如NPR[29],在JPEG压缩和尺寸缩放后性能下降明显。此外,大部分方法在下采样操作下的鲁棒性弱于上采样,原因在于下采样过程会损失更多的图像细节。通过高质量扩散模型生成具有类似后处理干扰的困难样本,使模型提前适应干扰,并借助对比损失函数引导模型聚焦真实与生成图像间不可消除的扩散伪影差异,从而过滤易受后处理影响的表层特征,如区分JPEG压缩块效应与扩散伪影。在高斯噪声鲁棒性实验中,本文方法平均准确率达到90.6%,尽管随着噪声标准差增大性能有所下降,但整体仍表现出较优的检测效果。该实验不仅验证了CLDA在真实场景的实用性,更证明聚焦扩散固有伪影、采用难样本预训练加上对比学习的策略,能够有效突破后处理干扰带来的瓶颈,为后续鲁棒性优化提供了重要思路。

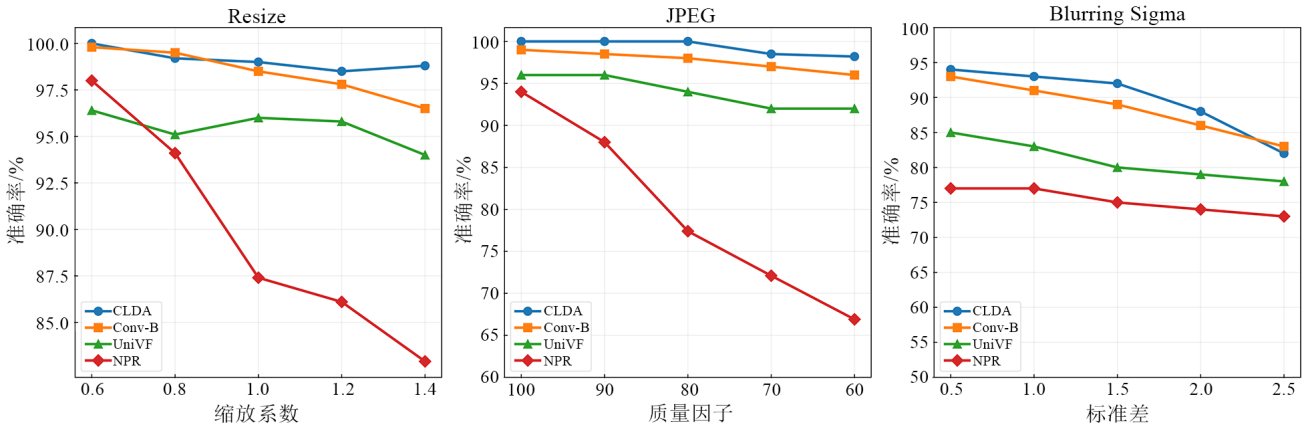


图7 GenImage子集SDv1.4上调整大小、JPEG压缩和高斯噪声的鲁棒性评估

Figure 7 Robustness evaluation of resizing, JPEG compression, and Gaussian noise on the GenImage subset SDv1.4

3.4 消融实验

在消融研究中,本文重点评估了两个关键设计对检测性能的影响:基于潜在空间边界的对比损失函数,以及伪影增强模块的引入。为保持实验一致性,所有模型均在GenImage数据集的SDv1.4子集上训练,并在完整GenImage测试集上评估。训练完成后,本文在GenImage数据集上对相关模型进行了性能测试。实验结果如表4所示,基线模型初始平均准确率为67.2%。引入伪影增强模块后,准确率显著提升至73.4%;进一步加入域损失后,准确率再提高3.9个百分点;最终采用基于潜在空间边界的对比学习策略,准确率进一步上升5.3个百分点,整体达到83.6%。实验结果表明,通过对比学习与边界损失相结合,模型能够有效捕捉扩散模型生成图像中的特异性伪影,并显著抑制对语义特征的过拟合。消融研究充分验证了本文所提扩散伪影对比学习框架的有效性及各模块的协同作用。

表4 消融实验性能测试 单位:%

Table 4 Performance test of ablation experiment unit: %

w/o 增强模块	w/o 域损失	w/o 边界对比损失	AVG. ACC
×	×	×	67.2
√	×	×	73.4
√	√	×	77.3
√	√	√	**83.6**

注:加粗为对比最优结果。

3.5 可视化分析

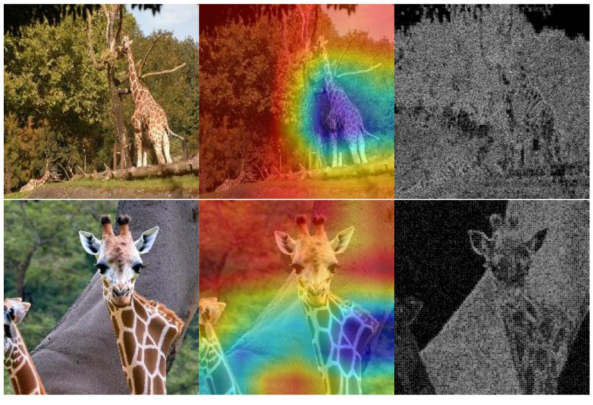
为探究扩散伪影对比学习中所捕捉的伪影特性,本文进一步通过类激活图(Class Activation Map,CAM)对模型的关注区域进行可视化分析。类激活图能够清晰揭示网络在分类决策过程中所依赖的关键图像区域:特征图中权重越高的区域,对最终分类结果的贡献越大,其重要性也越突出。

CLDA方法的核心思想是通过真实图像与生成图像的对比学习,识别最具判别力的通用伪影特征。图8展示了真实图像与生成图像的示例及其对应的SRM(隐写分析丰富模型)特征输出。从图8可知,真实图像对应的模型关注区域更集中于自然结构细节,而生成图像在关注区域分布及噪声指纹特征方面均表现出明显差异。为量化CAM关注区域与物理线索之间的关联性,本文通过频域分析提取图像的高频残差,并计算关注区域内的残差能量。其中,频域残差定义为高频分量与原始频域分量之间的幅值差异,计算公式如下:

$$\mathrm{Res}(x,y)=\left|F_{\mathrm{high}}(x,y)-F(x,y)\right| \tag{13}$$

其中,(x, y)为频域图像坐标;F为经过傅里叶变换得到的频域图像;F_{high}是经过滤波后的高频分量,对CAM关注区域残差值取平方求和,得到残差能量。以下图为例,真实图像的局部细节残差能量占整图的69.4%,而虚假图像占比仅27.7%,说明模型关注的虚假纹理区域高频异常强度低且分散,缺乏物理依据。该差异验证了真实与虚假样本在物理线索上的本质区别。

本文方法能够增强模型对扩散生成图像中伪影特征的感知能力,使其能够更准确地定位并理解图像中的伪造痕迹。这种增强的感知机制使模型能够敏锐捕捉生成图像与真实图像在细节伪影方面的差异,从而提升检测任务的准确性与鲁棒性。



注:上图为真实图像,下图为生成图像。

图8　CAM和SRM可视化图片

Figure 8　CAM and SRM visualization images

4 结论

本文提出了一种基于扩散伪影的对比学习框架,并设计了相应的伪影增强模块。该模块通过潜在空间增强技术扩展伪造图像的特征表示范围,引导模型学习更具泛化性的潜在空间边界,从而提升检测任务的鲁棒性,并促进模型捕捉不同伪造图像中的共性伪影特征。为进一步增强模型判别能力,本文引入域损失机制,将每种伪造类型与真实图像视为独立域,驱动编码器学习具有域区分能力的特征表示,以提高对多类别图像的识别性能。同时,本文设计了基于潜在空间边界的对比损失函数,利用反演图像中固有的模型指纹作为关键线索,引导检测器精准学习真实图像与生成图像之间的细微差异。CLDA框架有效提升了现有检测器对扩散生成图像的识别精度与泛化能力。实验结果表明,在GenImage数据集上,本文方法相比当前最佳结果将检测准确率提升了1.1个百分点;在DRCT-2M数据集上,准确率提升达4.8个百分点,充分体现出其优异的跨数据泛化性能。

参考文献

[1] 惠康华, 闫建青, 高思华, 等. 基于特征融合的轻量级新残差人脸识别方法[J]. 电子学报, 2024, 52(3): 937-944.
Hui Kanghua, Yan Jianqing, Gao Sihua, et al. Lightweight new fesidual face recognition method based on feature fusion[J]. Acta Electronica Sinica, 2024, 52(3): 937-944. (in Chinese)

[2] Gu Fei, Dai Yunshu, Fei Jianwei, et al. Deepfake detection and localisation based on illumination inconsistency[J]. International Journal of Autonomous and Adaptive Communications Systems, 2024, 17(4): 352-368.

[3] 何琨, 佘计思, 张子君, 等. 基于引导扩散模型的自然对抗补丁生成方法[J]. 电子学报, 2024, 52(2): 564-573.
He Kun, She Jisi, Zhang Zijun, et al. A guided diffusion-based approach to natural adversarial patch generation[J]. Acta Electronica Sinica, 2024, 52(2): 564-573. (in Chinese)

[4] Yang Ling, Zhang Zhilong, Song Yang, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2024, 56(4): 105.

[5] Zhang Xu, Karaman S, Chang S F. Detecting and simulating artifacts in GAN fake images[C]//2019 IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE, 2019: 9035107.

[6] Juefei-Xu F, Wang Run, Huang Yihao, et al. Countering malicious DeepFakes: Survey, battleground, and horizon[J]. International Journal of Computer Vision, 2022, 130(7): 1678-1734.

[7] Wang Zhendong, Bao Jianmin, Zhou Wengang, et al. DIRE for diffusion-generated image detection[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 22388-22398.

[8] Zhu Mingjian, Chen Hanting, Yan Qiangyu, et al. GenImage: A million-scale benchmark for detecting AI-generated image[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2023: 3398.

[9] Chen Baoying, Zeng Jishen, Yang Jianquan, et al. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 7621-7639.

[10] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 574.

[11] Song Jiaming, Meng Chenlin, Ermon S. Denoising diffusion implicit models[C]//Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

[12] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2021: 672.

[13] Tan Chuangchuang, Tao Renshuai, Liu Huan, et al. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(7): 7184-7192.

[14] Zhao Shihao, Chen Dongdong, Chen Y C, et al. Uni-ControlNet: All-in-one control to text-to-image diffusion models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2023: 491.

[15] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10674-10685.

[16] Li Lingzhi, Bao Jianmin, Zhang Ting, et al. Face X-ray for more general face forgery detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5000-5009.

[17] Vasilcoiu A, Najdenkoska I, Geradts Z, et al. LATTE: Latent trajectory embedding for diffusion-generated image detection[PP/OL]. V2.arXiv (2025-09-29)[2025-10-10]. https://arXiv.org/abs/2507.03054.

[18] Wang Shengyu, Wang O, Zhang R, et al. CNN-generated images are surprisingly easy to spot... for now[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 8692-8701.

[19] 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. JMLR.org, 2020: 304.

[20] Yu Ning, Davis L, Fritz M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 7555-7565.

[21] 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. PMLR, 2021: 8748-8763.

[22] Ojha U, Li Yuheng, Lee Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 24480-24489.

[23] Li Xinghan, Yu Yue, Song Xue, et al. Revealing the implicit noise-based imprint of generative models[PP/OL]. V2.arXiv (2025-11-16)[2025-10-10]. https://arXiv.org/abs/2503.09314.

[24] Liu Bo, Yang Fan, Bi Xiuli, et al. Detecting generated images by real images[C]//Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 95-110.

[25] Guarnera L, Giudice O, Battiato S. Level up the deepfake detection: A method to effectively discriminate images generated by GAN architectures and diffusion models[M]//Arai K. Intelligent systems and applications. Cham: Springer, 2024: 615-625.

[26] Guo Xiao, Liu Xiaohong, Ren Zhiyuan, et al. Hierarchical fine-grained image forgery detection and localization[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 3155-3165.

[27] Qian Yuyang, Yin Guojun, Sheng Lu, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//Proceedings of the 16th European Con-

ference on Computer Vision. Heidelberg: Springer, 2020: 86-103.

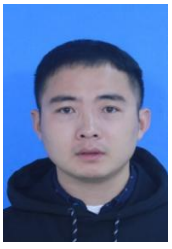
[28] Liu Zhengzhe, Qi Xiaojuan, Torr P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 8057-8066.

[29] Tan Chuangchuang, Liu Huan, Zhao Yao, et al. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 28130-28139.

[30] Luo Yunpeng, Du Junlong, Yan Ke, et al. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 17006-17015.

[31] Liu Zhuang, Mao Hanzi, Wu Chaoyuan, et al. A ConvNet for the 2020s[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11966-11976.

作者简介



袁程胜　男,1989年5月出生于江苏省连云港市。现为南京信息工程大学计算机学院、网络空间安全学院副教授。主要研究方向为人工智能安全。

E-mail: yuancs@nuist.edu.cn

陈金瑞　男,2001年11月出生于江苏省淮安市。现为南京信息工程大学计算机学院、网络空间安全学院硕士研究生。主要研究方向为人工智能安全。

E-mail: 202312490570@nuist.edu.cn

曹　燚　男,1994年7月出生于江苏省宿迁市。现为无锡学院网络安全与信息化学院副教授。主要研究方向为信息隐藏。

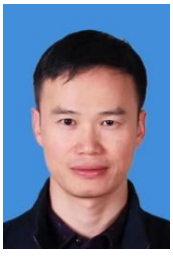
E-mail: caoyi@cwxu.edu.cn

刘庆程　男,1987年12月出生于安徽省滁州市。现为南京信息工程大学计算机学院、网络空间安全学院博士研究生。主要研究方向为多媒体内容安全、电力大数据挖掘与分析。

E-mail: qingcheng_liu@nuist.edu.cn

周志立　男,1984年6月出生于湖北省武穴市。现为广州大学人工智能研究院教授。主要研究方向为信息安全、大模型应用、智慧医疗。

E-mail: zhou_zhili@gzhu.edu.cn



付章杰　男,1983年3月出生于河南省南阳市。现为南京信息工程大学计算机学院、网络空间安全学院教授、院长。主要研究方向为信息安全。中国电子学会会员编号:E190023728M。

E-mail: wwwfzj@126.com