Computer Science and Application 计算机科学与应用, 2026, 16(7), 37-46
Published Online July 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.167239

Hans汉斯

深度伪造人脸图像检测方法研究进展：从强监督到无监督学习

魏雅霓，王丹琳*

北京警察学院刑事科学技术系，北京

收稿日期：2026年6月6日；录用日期：2026年7月7日；发布日期：2026年7月14日

摘　要

随着生成式人工智能快速发展，深度伪造人脸图像的真实感不断提升，对金融安全、司法取证和社会信任构成威胁，深度伪造人脸图像检测因此成为数字内容安全领域的重要研究方向。本文首先分析自编码器、生成对抗网络、语义可控生成和扩散模型等生成技术演进对检测线索变化的影响；其次按照训练过程中对标注信息的依赖程度，将现有检测方法归纳为强监督检测、弱监督与半监督检测、无监督检测三类，并总结其技术特点与局限。综述表明，相关研究正由显性伪影识别转向隐性特征建模，由闭集二分类转向跨域泛化与开放场景适应。未来需重点提升隐性特征挖掘、低标注学习、跨域泛化、复杂扰动鲁棒性和可解释评测能力，并关注视觉基础模型、多模态一致性检测与主动防御技术的发展。

关键词

深度伪造检测，人脸图像，生成对抗网络，无监督学习，跨域自适应

Advances in Deepfake Facial Image Detection Methods: From Fully Supervised to Unsupervised Learning

Yani Wei, Danlin Wang*

Department of Criminal Science and Technology, Beijing Police College, Beijing

Received: June 6, 2026; accepted: July 7, 2026; published: July 14, 2026

Abstract

With the rapid development of generative artificial intelligence, deepfake facial images have become

*通讯作者。

文章引用: 魏雅霓, 王丹琳. 深度伪造人脸图像检测方法研究进展: 从强监督到无监督学习[J]. 计算机科学与应用, 2026, 16(7): 37-46. DOI: 10.12677/csa.2026.167239

increasingly realistic, posing threats to financial security, judicial forensics, and social trust. Deepfake facial image detection has therefore become an important research direction in digital content security. This paper first analyzes how the evolution of generative techniques, including autoencoders, generative adversarial networks, semantically controllable generation, and diffusion models, has changed detectable forgery traces. Then, according to the dependence on annotation information during training, existing detection methods are categorized into three groups: fully supervised detection, weakly supervised and semi-supervised detection, and unsupervised detection, with their technical characteristics and limitations summarized. This review shows that related research is shifting from explicit artifact recognition to implicit feature modeling, and from closed-set binary classification to cross-domain generalization and open-scenario adaptation. Future research should focus on improving implicit feature mining, low-annotation learning, cross-domain generalization, robustness against complex perturbations, and interpretable evaluation, while also paying attention to the development of vision foundation models, multimodal consistency detection, and active defense techniques.

Keywords

Deepfake Detection, Facial Images, Generative Adversarial Networks, Unsupervised Learning, Cross-Domain Adaptation

1. 引言

生成式人工智能技术的快速发展推动深度伪造人脸技术不断成熟。深度伪造通常依托自编码器、生成对抗网络、扩散模型等深度生成架构，对人脸图像进行身份替换、属性编辑、表情驱动与内容合成，能够生成具有较高真实感的伪造人脸图像。以 ZAO、Reface 等为代表的消费级应用降低了人脸替换和人脸编辑的技术门槛，深度伪造从专业化制作快速扩展至普通用户。技术易用性提升的同时，滥用风险也急剧扩大，虚假内容传播、身份伪造、舆论操纵等问题频发，对社会安全、个人权益、司法取证和公众信息信任体系构成严峻挑战。

在此背景下，深度伪造人脸图像检测成为数字内容安全和图像取证领域的重要研究方向。现有检测方法通常通过学习真实图像与伪造图像之间的差异，实现对伪造人脸图像的识别与判别。早期检测模型主要依赖边界融合异常、纹理模糊、颜色失配等显性视觉伪影，在自编码器式换脸和早期生成模型场景下能够取得一定效果。然而，随着生成对抗网络、语义解耦建模和扩散模型的发展，伪造人脸图像在身份一致性、表情自然度、光照协调性和局部纹理细节等方面不断提升，传统显性伪影逐渐弱化，检测模型面临跨生成器、跨数据集、跨压缩质量和跨应用场景泛化能力不足的问题。同时，大规模高质量标注数据获取困难，真实传播环境中的裁剪、压缩、转码和滤镜处理也会进一步削弱模型对细粒度伪造痕迹的捕捉能力。因此，深度伪造人脸图像检测技术正在由依赖有标签样本的强监督学习，逐步向弱监督、半监督、无监督学习以及开放场景泛化等方向发展。

本文聚焦深度伪造人脸图像检测问题，首先，概述深度伪造人脸生成技术的发展脉络，其次按照监督范式的变化，总结从传统强监督学习到弱监督、半监督和无监督检测的技术发展脉络。最后，围绕高保真生成条件下隐性伪造特征挖掘、开放生成源下跨域泛化、低标注与无标签学习、复杂扰动环境下鲁棒性以及可解释评测等问题，分析深度伪造人脸图像检测面临的挑战与未来发展方向。本文旨在为相关研究者理解深度伪造人脸图像检测技术演进规律、优化检测模型设计和推动实际应用提供参考。

2. 深度伪造人脸图像生成技术发展

2017 年，Reddit 网站上一个名为“deepfakes”的用户利用 TensorFlow、Keras 等深度学习框架，将自编码器模型与人脸替换流程相结合，制作并传播了早期深度伪造人脸内容，推动了深度伪造人脸技术的扩散[1]。深度伪造人脸生成技术的演进，直接改变了伪造图像中的可检测痕迹，也推动检测方法不断从显性伪影识别走向隐性特征建模、跨域泛化与无监督异常检测。早期深度伪造人脸生成主要以自编码器重建和面部区域替换为核心，通过学习源身份与目标身份的人脸隐空间表示，实现身份区域的重建、替换与后处理融合。受模型表达能力、训练数据规模和后处理效果限制，该阶段伪造图像常伴随边界不连续、纹理模糊、颜色失配等低层视觉伪影，为早期强监督检测模型提供了较为明确的判别依据。随后，生成对抗网络、身份保持约束和多尺度特征融合等机制被引入人脸替换和人脸编辑任务，伪造图像的纹理清晰度、身份一致性和视觉真实感明显提升，传统局部伪影特征的稳定性逐渐下降。随着语义解耦、三维结构约束和条件控制方法的发展，深度伪造人脸生成进一步从单一身份替换扩展到表情驱动、姿态控制、属性编辑等复杂任务，伪造痕迹也由边界、纹理等低层视觉异常转向身份一致性、几何结构一致性、光照协调性等更隐蔽层面。近年来，扩散模型及预训练生成模型进一步提升了人脸图像生成的开放性和高保真程度。与早期自编码器或 GAN 模型生成图像中较稳定的边界融合痕迹、上采样纹理异常和频域伪影不同，扩散模型通过逐步去噪过程生成图像，能够在局部纹理、噪声分布和高频细节上更接近真实图像分布。这使得真实图像与伪造图像在低层像素统计、频域分布和噪声残差上的差异进一步缩小，传统检测器依赖的固定伪影模式不再稳定存在。同时，扩散模型具有较强的开放生成能力和条件控制能力，不同提示词、控制条件、采样策略和后处理方式都可能改变生成图像的局部统计特征。由此导致伪造痕迹呈现出更强的非固定性和跨模型差异，检测模型如果仍然依赖某一生成器或某一数据集中的特定频率模式，容易在面对新型扩散模型生成图像时出现性能下降，使检测任务面临更强的跨生成器、跨数据集和跨应用场景泛化挑战。为更清晰地呈现不同阶段生成技术特征、代表性方法及其典型伪造痕迹，本文对深度伪造人脸图像生成技术的发展阶段进行归纳，如表 1 所示。

Table 1. Evolution of deepfake facial image generation technologies and typical forgery traces
表 1. 深度伪造人脸图像生成技术演进及典型伪造痕迹

生成阶段	主要技术特征	代表性方法	开源工具/商业化软件	典型伪造痕迹
自编码器重建与人脸替换阶段	以自编码器重建和面部区域替换为核心，通过学习源身份与目标身份的隐空间表示完成人脸替换	DeepFaceLab [2]	FakeApp [3] Faceswap [4] DeepFaceLab [2]	边界融合不自然、纹理模糊、颜色失配、局部细节缺失较明显
对抗生成与真实感增强阶段	引入生成对抗网络、身份保持约束、注意力机制和图像增强策略，提高人脸替换的清晰度和真实感	FaceShifter [5] FSGAN [6] SimSwap [7]	faceswap-GAN [8] ZAO [9] FaceApp [10]	显性边界伪影减少，纹理真实感增强，局部缺陷更加细微
语义解耦与结构约束的可控生成阶段	对身份、表情、姿态、光照和运动等因素进行解耦建模，实现更可控的人脸编辑、表情驱动	DiscoFaceGAN [11] StyleRig [12] PIRenderer [13]	Snapchat Face Swap Component [14] Reface [15]	低层伪影进一步弱化，伪造异常更多体现为几何结构、身份保持不稳定等方面
扩散模型驱动的开放式高保真生成阶段	通过扩散模型、预训练生成模型和条件控制机制实现更高保真、更开放的人脸生成与编辑	DiffFace [16] DiffSwap [17] DiffusionRig [18] Face-Adapter [19]	FaceFusion [20]	传统边界、纹理和频域伪影显著减弱，伪造样本更接近真实图像分布

3. 深度伪造人脸图像检测技术分类与演进

随着深度伪造人脸图像生成技术不断升级，伪造痕迹越来越隐蔽，检测方法也从早期有标签样本的二分类判别，逐步扩展到多特征融合、低标注表征学习、无监督异常检测和开放场景下的跨域泛化。围绕“从强监督到无监督学习”的技术演进主线，本文将现有深度伪造人脸图像检测方法划分为基于有标签样本的强监督检测、面向低标注场景的弱监督与半监督检测方法、面向未知伪造的无监督检测三类方法，其监督范式演进关系如图 1 所示。

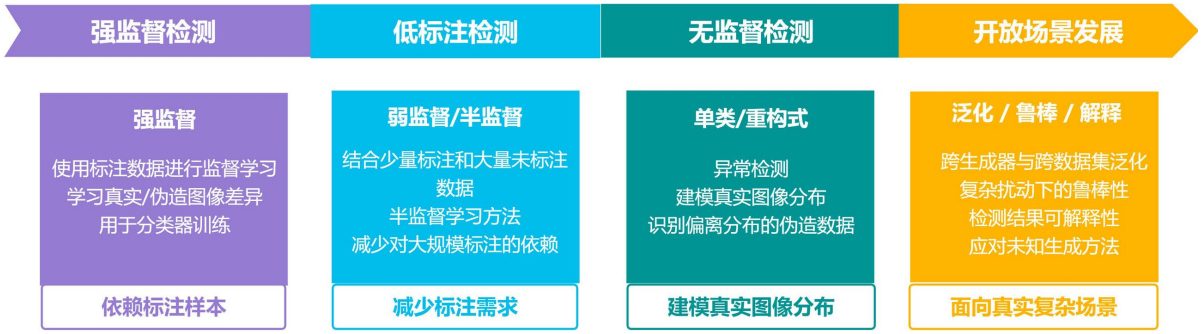

Figure 1. Framework of supervision paradigm evolution in deepfake facial image detection methods
图 1. 深度伪造人脸图像检测方法的监督范式演进框架

3.1. 基于有标签样本的强监督检测方法

基于有标签样本的强监督检测方法是深度伪造人脸图像检测中发展较早、应用较广的技术范式。该类方法通常将检测任务建模为真实图像与伪造图像之间的二分类问题，利用人工标注样本学习两类图像在空间纹理、频域统计、噪声残差、融合边界和结构一致性等方面的差异。根据特征建模对象的不同，强监督检测方法可进一步分为空间域伪影检测、频域统计与图像取证特征检测、多特征融合与结构一致性检测等方向。

3.1.1. 空间域伪影检测方法

空间域伪影检测方法主要从图像纹理、边界融合、颜色分布和局部结构异常等角度识别伪造痕迹。早期自编码器人脸和初期 GAN 伪造图像常有边界融合不自然、颜色失配、纹理模糊等显性伪影，因此基于卷积神经网络的空间域检测方法成为早期主流。MesoNet [21]是空间域伪影检测的代表性方法之一，该方法采用结构紧凑的浅层卷积网络，重点提取人脸局部纹理异常，具有模型规模小、检测效率较高等特点。由于其网络深度和特征表达能力有限，面对高质量伪造样本时检测性能容易下降。Xception [22]基于深度可分离卷积构建，通过多尺度特征提取捕捉全局纹理与空间篡改痕迹，具有较强的图像特征提取能力。相较于浅层 CNN，Xception 能够学习更丰富的空间纹理和局部结构特征，但其性能仍然依赖训练数据的伪造类型、压缩质量和分布特征。提出的 Noiseprint 方法则通过增强相机模型相关伪迹，为利用成像指纹识别伪造图像提供了思路。空间域方法在显性伪影较明显的场景中较为有效，但其性能容易受到伪造类型、压缩质量和训练数据分布的影响。

3.1.2. 频域统计与图像取证特征检测方法

随着生成模型真实感不断提升，单纯依赖 RGB 空间纹理特征的检测方法逐渐失效。检测方法开始从单纯的空间纹理，扩展到频域异常、局部纹理统计、颜色空间差异、噪声残差、相机成像痕迹等多维度线索。F^3-Net [23]是频域伪影检测的代表性方法之一，其通过离散余弦变换引入频域表征，并设计频率分

解和局部频率统计两个互补分支，以捕捉高质量伪造图像中肉眼难以察觉的频率异常。除频域特征外，Xu 等人[24]利用梯度、标准差、灰度共生矩阵和小波变换等纹理特征进行检测；Nataraj 等人[25]基于 RGB 通道像素共生矩阵刻画生成图像的局部统计关系；Barni 等人[26]进一步利用跨波段共生矩阵分析 GAN 生成人脸图像的颜色通道相关性；Cozzolino 和 Verdoliva [27]提出的 Noiseprint 方法则通过增强相机模型相关伪迹，为利用成像指纹识别伪造图像提供了思路。此类方法能够从空间纹理之外提供补充判别信息，但仍可能受到压缩、分辨率变化、平台转码和不同生成器频率模式差异的影响。

3.1.3. 多特征融合与结构一致性检测方法

多特征融合与结构一致性检测方法旨在综合利用局部纹理、边界痕迹、频域异常、身份保持、几何结构和语义一致性等多类线索，以提升模型对复杂伪造样本的适应能力。在人脸图像检测中，伪造痕迹不仅可能表现为局部区域的纹理异常，也可能体现为人脸边界、五官结构、光照分布、身份特征和背景关系之间的不协调。Face X-Ray [28]从人脸伪造过程中常见的图像融合操作出发，构建用于揭示不同来源区域融合边界的灰度表示，从而提升模型对未知伪造类型的检测能力。Multi-Attentional Deepfake Detection [29]则将深度伪造检测视为细粒度分类问题，通过多个空间注意力头引导模型关注不同局部区域，并结合纹理增强模块捕捉细微伪造痕迹。此类方法说明，在高质量伪造场景下，检测模型不能仅依赖整图分类特征，还需要关注局部异常区域与整体人脸结构之间的关系。

总体来看，基于有标签样本的强监督检测方法在已知伪造类型和标注数据充足的条件下能够取得较好效果，是深度伪造人脸图像检测的重要基础。但该类方法通常依赖大规模高质量标注数据，且容易学习到特定数据集、特定生成器或特定压缩条件下的伪影模式。当面对未知生成方法、复杂后处理或跨数据集测试时，模型泛化能力仍可能下降。这一局限推动了弱监督、半监督、无监督检测方法的发展，也促进了自监督表征学习在深度伪造检测中的应用。

3.2. 面向低标注场景的弱监督与半监督检测方法

随着深度伪造生成技术快速迭代，新型伪造样本不断出现，传统强监督检测方法对大规模精确标注数据的依赖逐渐成为限制其实际应用的重要因素。尤其在真实应用场景中，伪造样本来源复杂、生成方法更新迅速，人工标注难以及时覆盖新型伪造类型。为降低对精确人工标注的依赖，弱监督、半监督检测方法逐渐受到关注，自监督表征学习也被用于提升低标注条件下模型的特征表达能力和泛化能力。

弱监督检测主要利用低成本、不完整或不完全可靠的监督信息学习判别特征，从而减少对像素级伪造区域标注和大量精确样本标签的依赖。半监督检测则通常结合少量有标签样本和大量无标签样本，通过一致性约束、伪标签生成、特征对齐等策略扩大训练数据覆盖范围。对于深度伪造人脸图像检测而言，这类方法能够缓解新型伪造样本标注滞后、目标域样本缺少标签以及跨生成器数据分布差异等问题。DPGNet [30]是该方向的代表性方法之一。该方法面向未知生成源下的大规模无标注人脸伪造数据利用问题，采用标注源域与无标注目标域联合训练框架，通过文本引导跨域对齐和课程驱动伪标签生成机制，缓解不同生成器之间的域差异，并提高模型对未知伪造来源的适应能力。此类方法的优势在于能够充分利用无标签目标域数据，提升模型在新场景中的迁移能力；但其性能也容易受到伪标签质量、目标域数据复杂性和源域－目标域分布差异的影响。当伪标签错误不断累积时，模型可能学习到不稳定甚至错误的判别边界。

在低标注检测框架中，自监督学习作为一种通用表征学习工具发挥作用。自监督学习通过对比学习、掩码重建、图像恢复等方式从数据自身构造监督信号，学习具有泛化能力的视觉表征。在少量标注样本条件下，自监督预训练特征可以为弱监督或半监督检测提供更稳定的特征初始化，降低模型对特定生成器伪影的过度依赖；在无标签目标域中，自监督表征也可以与伪标签生成、域对齐和一致性约束结合，

提高模型对未知伪造来源的适应能力。例如，SLADD [31]是自监督学习用于提升深度伪造检测泛化能力的代表性研究之一。该方法通过自监督学习构造对抗样本，使模型在训练过程中接触更加多样的伪造扰动，从而增强其对不同伪造类型的敏感性和跨方法泛化能力。与仅依赖特定伪造数据训练的强监督方法相比，自监督表征学习有助于降低模型对某一类生成器伪影的过度依赖，使其学习更一般化的判别线索。然而，自监督学习预训练任务与最终真假判别目标之间可能存在差异，模型学习到的表征未必都能直接服务于深度伪造检测。

总体来看，弱监督和半监督方法主要通过低成本标签、伪标签和无标签目标域数据利用来降低标注依赖；自监督学习则通过通用表征学习为低标注检测提供特征支撑。低成本监督信息利用与自监督表征学习相结合，有助于推动深度伪造检测从依赖精确标注的闭集分类，向低标注、跨生成源和开放场景泛化方向发展。该类方法能够缓解标注成本高、伪造样本更新快和目标域无标签等问题，但仍面临伪标签噪声传播、自监督任务与检测目标不完全一致、跨域表征稳定性不足等挑战。

3.3. 基于无标签样本的无监督检测方法

相较于强监督检测和低标注检测方法，无监督检测进一步减少了对伪造样本及其标签的依赖。该类方法不再依赖已知伪造样本进行真假二分类训练，而是通过建模真实图像分布、学习正常样本特征，将偏离真实分布的图像识别为异常或潜在伪造。由于无监督检测不要求训练阶段覆盖具体伪造类型，因此在应对新型生成模型和未知伪造样本方面具有一定潜力。

3.3.1. 单类分类与真实分布建模方法

单类分类方法通常仅利用真实人脸样本进行训练，通过学习真实样本在特征空间中的紧凑分布，将远离该分布的样本判定为异常。Deep SVDD [32]等深度单类学习方法为该方向提供了重要思路。其核心思想是将正常样本映射到特征空间中的紧凑区域，并通过度量待检测样本与该区域之间的距离判断其是否异常。将这一思想引入深度伪造人脸图像检测时，模型可通过学习真实人脸图像在纹理、结构、频域或深层表征上的正常分布，识别那些明显偏离该分布的伪造样本。然而，单类分类方法也面临真实分布复杂、异常边界难以设定等问题。真实人脸图像并不是简单的单峰分布，不同身份、姿态和成像条件可能导致较大类内差异。

3.3.2. 重构式异常检测方法

异常检测方法则通常利用自编码器、变分自编码器或生成模型建模真实人脸分布，并根据重构误差、潜在空间偏移或特征分布差异判断图像是否异常。例如，OC-FakeDect [33]是该方向的代表性方法之一。该方法将深度伪造检测建模为单类异常检测问题，利用 one-class VAE 仅在真实人脸图像上学习正常分布，并将 deepfake 图像视为偏离真实分布的异常样本。与依赖已知伪造类型进行监督训练的方法相比，重构式异常检测更强调对真实人脸分布的建模，因此在检测未知伪造样本方面具有一定优势。然而，重构式方法同样存在局限。一方面，高质量伪造图像可能与真实图像高度相似，导致重构误差并不明显；另一方面，真实图像中的极端姿态、特殊光照、低分辨率或压缩失真也可能造成较大重构误差，从而引发误判。

总体来看，基于无标签样本的无监督检测方法体现了深度伪造检测从“学习已知伪造痕迹”向“建模真实人脸分布”的转变。该类方法能够降低对伪造样本标签的依赖，并为未知伪造检测提供思路，但其性能仍受真实样本分布复杂性、异常阈值设定和跨数据集泛化能力的限制。未来，无监督检测仍需与自监督表征学习、鲁棒特征建模和开放场景泛化方法结合，才能更好适应高保真生成条件下的深度伪造人脸图像检测需求。

4. 当前挑战与未来方向

随着深度伪造生成技术由早期自编码器、GAN 进一步发展至语义可控生成和扩散模型阶段，伪造图像的真实感、细节一致性和生成开放性持续增强。相应地，深度伪造人脸图像检测也面临由显性伪影识别向隐性特征建模、由闭集分类向开放场景泛化、由依赖精确标注向低标注和无标签学习转变的多重挑战。

4.1. 高保真生成条件下的隐性伪造特征挖掘

高保真生成模型显著削弱了传统检测方法依赖的边界错位、纹理模糊、色彩失衡等显性伪影，使单纯依赖局部瑕疵的检测方法逐渐失效。未来研究需要从“捕捉显性伪影”转向“挖掘隐性判别特征”，重点关注频域统计异常、噪声残差、生成模型残留模式、潜在特征分布偏移以及局部–全局结构一致性等更深层线索，以提升模型对高保真伪造图像的识别能力。

4.2. 开放生成源下的跨域泛化能力提升

现有检测方法大多建立在闭集分类假设之上，训练集通常只能覆盖有限的伪造类型和生成器分布。然而，新型生成模型、开源工具和商业化人脸工具不断涌现，测试样本往往来自训练阶段未见过的伪造源，导致模型在跨生成器、跨数据集和跨场景测试中性能下降。针对这一问题，FeatureTransfer [34]、UDA-Fake [35]等研究已从无监督域适应和跨域特征对齐角度探索了提升跨数据集检测稳定性的路径，未来应进一步发展开放集检测、跨域泛化和域自适应方法，使模型学习与具体生成器弱相关的通用伪造表征，而不是依赖某一类生成模型的特定伪影。

4.3. 低标注与无标签条件下的通用表征学习

深度伪造技术迭代速度快，人工标注难以及时覆盖不断出现的新型伪造样本，传统依赖大规模精确标签的强监督检测模式面临局限。弱监督、半监督和无监督学习为降低标注依赖提供了新的路径，但仍存在监督信号可靠性不足、伪标签噪声累积、自监督任务与检测目标不一致、真实分布边界难以确定等问题。未来研究需要设计更贴合伪造检测任务的预训练目标、伪标签筛选机制和异常判别策略，提高模型在低标注和无标签条件下区分真实分布变化与伪造异常的能力。

4.4. 复杂扰动与反检测环境下的鲁棒性增强

真实传播环境中的图像往往会经历压缩、裁剪、缩放、滤镜、截图、平台转码等后处理操作，这些扰动会破坏模型依赖的细粒度伪影特征。同时，针对检测模型的对抗扰动、伪影抹除和反检测优化等手段也会进一步削弱模型性能。因此，未来检测方法需要同时提升对自然扰动和主动规避攻击的鲁棒性，可通过数据增强、域随机化、对抗训练、鲁棒特征学习和模型不确定性估计等方式，提高模型在复杂环境下的稳定性。

4.5. 可解释建模与可复现评测

现有深度伪造检测模型多以端到端分类为主，虽然在特定数据集上能够取得较高性能，但其判别依据往往不够透明，难以判断模型究竟学习到真实伪造线索，还是依赖数据集偏差、压缩痕迹或背景捷径完成分类。未来研究需要加强伪造区域定位、特征归因分析和可解释可视化方法，使模型不仅能够判断“是否伪造”，还能够说明“依据何在”。同时，当前不同研究在数据集划分、压缩等级、未知生成器设置和评价指标方面存在差异，导致结果可比性不足。构建统一的跨数据集、跨生成器、跨扰动条件评测协议和可复现基准，将有助于更加客观地评价不同检测方法的泛化能力与实际适用性。

4.6. 视觉基础模型与多模态一致性检测

随着视觉基础模型的发展，深度伪造检测开始从依赖专门训练的单一分类模型，转向利用大规模预训练表征提升跨域泛化能力。以 CLIP [36]、DINO [37]等为代表的视觉基础模型，通过大规模图像–文本对齐或自监督学习获得较强的通用视觉表征，为低标注、跨数据集和未知生成源检测提供了新的技术路径。但视觉基础模型主要面向通用语义表征学习，其关注重点未必与图像取证中的细粒度伪造痕迹完全一致，如何兼顾通用语义理解能力与取证敏感性，仍是后续研究需要解决的问题。除单一图像检测外，多模态深度伪造检测也是重要发展方向。现实场景中的深度伪造内容往往以视频、音频、文本等多种形式组合传播，仅依赖单帧人脸图像可能难以发现复杂伪造。音视频多模态检测可以从人脸动作、唇形变化、语音内容、说话人身份和情绪表达等多个维度分析一致性。因此，未来深度伪造检测需要从静态图像判别进一步扩展到音视频协同分析、时序一致性建模和跨模态语义一致性检测。

4.7. 从被动检测到主动防御

现有深度伪造检测方法大多属于事后被动检测，即在图像或视频生成、传播之后判断其是否存在伪造痕迹。随着生成模型真实感不断提升，仅依赖被动检测可能难以满足复杂网络环境中的真实性验证需求。数字水印、生成内容标识和来源追溯技术能够在内容生成、编辑和传播过程中记录媒体来源及修改历史，为后续真实性验证提供依据。其中，C2PA [38]等内容来源与真实性标准尝试通过可验证元数据记录数字内容的来源、编辑过程和身份声明，从而为媒体可信度判断提供辅助信息。与传统检测模型相比，主动防御技术更强调在内容生成和传播链条中提前嵌入可验证信息，能够为深度伪造内容识别提供另一种技术路径。但主动防御技术仍面临平台支持程度不一、元数据可能丢失或被恶意移除、跨平台兼容性不足等问题。因此，未来更可行的发展方向并不是以主动防御完全替代被动检测，而是将深度伪造检测、数字水印、来源追溯和平台治理机制结合起来，形成多层次的深度伪造风险防控体系。

5. 结论

深度伪造人脸图像生成技术的持续演进，使伪造内容的真实感、开放性和隐蔽性不断增强，也推动检测方法由早期依赖显性伪影识别，逐步转向频域统计、噪声残差、结构一致性和深层表征等隐性特征建模。本文围绕“从强监督到无监督学习”的技术主线，对深度伪造人脸图像检测方法进行了梳理。总体来看，强监督方法在已知伪造类型和标注数据充分的条件下仍具有较好效果，但其泛化能力受数据分布和特定伪影模式限制；弱监督、半监督和无监督方法为降低标注依赖、适应未知伪造提供了新的思路，但仍面临伪标签噪声、表征稳定性不足和真实分布边界难以确定等问题。未来研究应进一步提升隐性特征挖掘、跨域泛化、复杂扰动鲁棒性和可解释评测能力，并结合视觉基础模型、多模态一致性检测、数字水印和来源追溯等技术，推动检测方法向更通用、稳健和可信的方向发展。

基金项目

北京警察学院大学生创新训练计划项目资助(2025 市创大学生项目 09)。

参考文献

[1] Cole, S. (2018) We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. Motherboard/Vice. https://www.vice.com/en/article/reddit-fake-porn-app-daisy-ridley/

[2] Perov, I., Gao, D., Chervoniy, N., *et al*. (2020) DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework. arXiv:2005.05535.

[3] Shahzad, H.F., Rustam, F., Flores, E.S., Luís Vidal Mazón, J., de la Torre Diez, I. and Ashraf, I. (2022) A Review of

Image Processing Techniques for Deepfakes. *Sensors*, **22**, Article 4556. https://doi.org/10.3390/s22124556

[4] Faceswap Team (2017) Faceswap. GitHub. https://github.com/deepfakes/faceswap

[5] Li, L., Bao, J., Yang, H., Chen, D. and Wen, F. (2020) Advancing High Fidelity Identity Swapping for Forgery Detection. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5074-5083. https://doi.org/10.1109/cvpr42600.2020.00512

[6] Nirkin, Y., Keller, Y. and Hassner, T. (2019) FSGAN: Subject Agnostic Face Swapping and Reenactment. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 7184-7193. https://doi.org/10.1109/iccv.2019.00728

[7] Chen, R., Chen, X., Ni, B. and Ge, Y. (2020) SimSwap: An Efficient Framework for High Fidelity Face Swapping. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 2003-2011. https://doi.org/10.1145/3394171.3413630

[8] Shaoanlu (2026) Faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN

[9] Tang, F.X. (2019) Chinese 'Deepfake' App Censured over Privacy Concerns. Sixth Tone. https://www.sixthtone.com/news/1004513

[10] FaceApp Technology Limited (2026) FaceApp: Perfect Face Editor. https://www.faceapp.com/

[11] Deng, Y., Yang, J., Chen, D., Wen, F. and Tong, X. (2020) Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5154-5163. https://doi.org/10.1109/cvpr42600.2020.00520

[12] Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H., Perez, P., *et al*. (2020) StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6142-6151. https://doi.org/10.1109/cvpr42600.2020.00618

[13] Ren, Y., Li, G., Chen, Y., Li, T.H. and Liu, S. (2021) PiRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 13759-13768. https://doi.org/10.1109/iccv48922.2021.01350

[14] Snap Inc. (2026) Face Swap Component. Snap for Developers. https://developers.snap.com/lens-studio/features/ar-tracking/face/face-swap

[15] https://reface.ai/

[16] Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., *et al*. (2025) Diffface: Diffusion-Based Face Swapping with Facial Guidance. *Pattern Recognition*, **163**, Article 111451. https://doi.org/10.1016/j.patcog.2025.111451

[17] Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J. and Lu, J. (2023) Diffswap: High-Fidelity and Controllable Face Swapping via 3d-Aware Masked Diffusion. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 8568-8577. https://doi.org/10.1109/cvpr52729.2023.00828

[18] Ding, Z., Zhang, X., Xia, Z., Jebe, L., Tu, Z. and Zhang, X. (2023) DiffusionRig: Learning Personalized Priors for Facial Appearance Editing. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 12736-12746. https://doi.org/10.1109/cvpr52729.2023.01225

[19] Han, Y., Zhu, J., He, K., Chen, X., Ge, Y., Li, W., *et al*. (2024) Face-Adapter for Pre-Trained Diffusion Models with fine-Grained ID and attribute Control. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G., Eds., *Lecture Notes in Computer Science*, Springer, 20-36. https://doi.org/10.1007/978-3-031-72973-7_2

[20] FaceFusion Team (2026) FaceFusion. GitHub. https://github.com/facefusion/facefusion

[21] Afchar, D., Nozick, V., Yamagishi, J. and Echizen, I. (2018) MesoNet: A Compact Facial Video Forgery Detection Network. 2018 *IEEE International Workshop on Information Forensics and Security* (*WIFS*), Hong Kong, China, 11-13 December 2018, 1-7. https://doi.org/10.1109/wifs.2018.8630761

[22] Chollet, F. (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1251-1258. https://doi.org/10.1109/cvpr.2017.195

[23] Qian, Y., Yin, G., Sheng, L., Chen, Z. and Shao, J. (2020) Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, JM., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 86-103. https://doi.org/10.1007/978-3-030-58610-2_6

[24] Xu, B., Liu, J., Liang, J., Lu, W. and Zhang, Y. (2021) Deepfake Videos Detection Based on Texture Features. *Computers, Materials & Continua*, **68**, 1375-1388. https://doi.org/10.32604/cmc.2021.016760

[25] Nataraj, L., Mohammed, T.M., Chandrasekaran, S., *et al*. (2019) Detecting GAN Generated Fake Images Using Co-Occurrence Matrices. arXiv:1903.06836.

[26] Barni, M., Kallas, K., Nowroozi, E. and Tondi, B. (2020) CNN Detection of GAN-Generated Face Images Based on

Cross-Band Co-Occurrences Analysis. 2020 *IEEE International Workshop on Information Forensics and Security* (*WIFS*), New York, 6-11 December 2020, 1-6. https://doi.org/10.1109/wifs49906.2020.9360905

[27] Cozzolino, D. and Verdoliva, L. (2020) Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, **15**, 144-159. https://doi.org/10.1109/tifs.2019.2916364

[28] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., *et al*. (2020) Face X-Ray for More General Face Forgery Detection. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5001-5010. https://doi.org/10.1109/cvpr42600.2020.00505

[29] Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D. and Yu, N. (2021) Multi-Attentional Deepfake Detection. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 2185-2194. https://doi.org/10.1109/cvpr46437.2021.00222

[30] Yang, Z., Tao, R., Zhang, C., *et al*. (2025) Leveraging Unlabeled Data from Unknown Sources via Dual-Path Guidance for Deepfake Face Detection. arXiv:2508.09022.

[31] Chen, L., Zhang, Y., Song, Y., Liu, L. and Wang, J. (2022) Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 18710-18719. https://doi.org/10.1109/cvpr52688.2022.01815

[32] Ruff, L., Vandermeulen, R., Goernitz, N., *et al*. (2018) Deep One-Class Classification. *International Conference on Machine Learning*, **80**, 4393-4402.

[33] Khalid, H. and Woo, S.S. (2020) OC-Fakedect: Classifying Deepfakes Using One-Class Variational Autoencoder. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 14-19 June 2020, 656-657. https://doi.org/10.1109/cvprw50498.2020.00336

[34] Chen, B. and Tan, S. (2021) Featuretransfer: Unsupervised Domain Adaptation for Cross-Domain Deepfake Detection. *Security and Communication Networks*, **2021**, Article ID: 9942754. https://doi.org/10.1155/2021/9942754

[35] Wang, Q., Wang, X., Liu, Z., Bai, N., Zhao, M. and Pang, S. (2026) Unsupervised Domain Adaptation-Based Cross-Type Deepfake Image Detection. *IEEE Transactions on Image Processing*, **35**, 4411-4424. https://doi.org/10.1109/tip.2026.3685445

[36] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning. *Proceedings of Machine Learning Research*, **139**, 8748-8763.

[37] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., *et al*. (2021) Emerging Properties in Self-Supervised Vision Transformers. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9650-9660. https://doi.org/10.1109/iccv48922.2021.00951

[38] Coalition for Content Provenance and Authenticity (2026) Content Credentials: C2PA Technical Specification, Version 2.2. https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html