Hans Journal of Data Mining 数据挖掘, 2026, 16(2), 60-71
Published Online April 2026 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2026.162006

Hans汉斯

基于混合防御架构的深度伪造检测与防御系统

姬懿轩，袁杨坤，甄博文，王喆宇，于　越，康晓凤

徐州工程学院信息工程学院，江苏　徐州

收稿日期：2026年3月8日；录用日期：2026年4月1日；发布日期：2026年4月10日

摘　要

深度伪造(Deepfake)技术的快速演进使高质量伪造内容泛滥，对个人隐私、舆论安全与社会治理构成严峻威胁。现有防御方法普遍存在检测手段单一、主动与被动防御割裂、跨数据集泛化能力不足等问题。为此，文章提出一种“主动防御 + 被动检测”的混合防御架构。在主动防御方面，提出梯度引导的区域自适应对抗扰动生成方法，结合人脸显著性图与梯度信息实现感知不可见的精准扰动分配，并设计基于DWT-DCT-SVD级联变换与K-means盲判决的鲁棒盲水印方案，支持“在线溯源 + 本地验真”双模式可信链。在被动检测方面，构建多源证据融合检测框架：以改进的双分支注意力增强ResNet-50为核心检测器，融合频域－纹理－几何传统特征分析层与视觉大模型语义推理层，并引入基于Dempster-Shafer证据理论的自适应融合决策机制替代固定权重方案，实现多源证据的不确定性建模与冲突消解。在FaceForensics++数据集上的实验表明，该系统性能良好，面对不同类型伪造Face2Face、FaceSwap、NeuralTextures检测准确率分别达到97.12%、95.58%和89.44%，优于现有主流方法。

关键词

深度伪造，混合防御架构，区域自适应对抗扰动，D-S证据融合，双分支注意力，盲水印

Deepfake Detection and Defense System Based on Hybrid Defense Architecture

Yixuan Ji, Yangkun Yuan, Bowen Zhen, Zheyu Wang, Yue Yu, Xiaofeng Kang

Department of Information Engineering, Xuzhou University of Technology, Xuzhou Jiangsu

Received: March 8, 2026; accepted: April 1, 2026; published: April 10, 2026

Abstract

The rapid evolution of Deepfake technology has led to the proliferation of high-quality forged content, posing a severe threat to personal privacy, public opinion security, and social governance. Existing defense methods generally have problems, such as single detection means, the separation of

文章引用: 姬懿轩, 袁杨坤, 甄博文, 王喆宇, 于越, 康晓凤. 基于混合防御架构的深度伪造检测与防御系统[J]. 数据挖掘, 2026, 16(2): 60-71. DOI: 10.12677/hjdm.2026.162006

active and passive defenses, and insufficient cross-dataset generalization ability. Therefore, this paper proposes a hybrid defense architecture of "active defense + passive detection". In terms of active defense, a method for generating gradient-guided region-adaptive adversarial perturbations is proposed. By combining face saliency maps and gradient information, precise perturbation allocation that is imperceptible to perception is achieved. A robust blind watermarking scheme based on the cascade transformation of DWT-DCT-SVD and K-means blind decision is designed to support the dual-mode trusted chain of "online traceability + local verification". In terms of passive detection, a multi-source evidence fusion detection framework is constructed: an improved dual-branch attention-enhanced ResNet-50 is used as the core detector, which integrates the traditional feature analysis layers of frequency domain-texture-geometry and the semantic reasoning layer of large vision models. An adaptive fusion decision mechanism based on Dempster-Shafer evidence theory is introduced to replace the fixed weight scheme, realizing the uncertainty modeling and conflict resolution of multi-source evidence. Experiments on the FaceForensics++ dataset show that the system has good performance. The detection accuracies for different types of forgeries, Face2Face, FaceSwap, and NeuralTextures, reach 97.12%, 95.58%, and 89.44%, respectively, which are better than existing mainstream methods.

Keywords

Deepfake, Hybrid Defense Architecture, Region-Adaptive Adversarial Perturbation, D-S Evidence Fusion, Dual-Branch Attention, Blind Watermark

1. 引言

近年来，以生成对抗网络(GAN) [1] [2]、自编码器[3]和扩散模型为代表的深度生成技术飞速迭代，深度伪造(Deepfake) [4]内容的逼真程度已接近人眼感知极限。此类技术在娱乐和影视中具有积极应用，但其滥用导致的虚假新闻传播、身份欺诈与名誉侵害等问题日趋严峻，对社会公信力和个人权益构成严重威胁[5]。因此，开发高效可靠的深度伪造检测与防御系统具有重要的现实意义和学术价值。

现有深度伪造防御研究可分为被动检测和主动防御两条路线[6]-[11]。被动检测方面，传统方法依赖频域特征、纹理特征和人脸几何特征进行判别，但面对高质量生成内容时鲁棒性不足；基于深度学习的方法虽然显著提升了检测精度，但单一模型的泛化能力有限，且缺乏可解释性。主动防御方面，现有方法主要包括数字水印和对抗扰动注入，但多数方案将两者独立设计，未能形成"防御–检测–溯源"的完整闭环。此外，鲜有工作将主动防御与被动检测在统一架构下进行协同设计，且多源检测结果的融合策略多采用固定权重，缺乏对证据不确定性和冲突的有效建模。

针对上述问题，本文提出融合主动防御与被动检测的混合防御架构(Hybrid Defense Architecture, HDA)，主要贡献如下：1) 提出梯度引导的区域自适应对抗扰动方法(Gradient-Guided Region-Adaptive Perturbation, GRAP)，利用人脸检测器输出的显著性图引导扰动在面部关键区域的精准分配，结合频域模式扰动与边界增强策略，在保持视觉质量的前提下显著提升对主流 Deepfake 生成管线的迁移干扰能力；2) 构建多源证据融合被动检测框架(Multi-Source Evidence Fusion, MSEF)，提出双分支注意力增强的 ResNet-50 (Dual-Branch Attention Enhanced ResNet, DBA-ResNet)，同时引入空间注意力与通道注意力分支捕获伪造痕迹的空间分布与通道响应特征，并采用基于 Dempster-Shafer 证据理论的自适应融合机制替代固定加权方案，

有效处理多源检测结果间的不确定性与冲突；3) 设计基于 DWT-DCT-SVD 级联变换的鲁棒盲水印方案，结合 K-means 聚类盲判决和双密码加密机制，支持“在线溯源 + 本地验真”双模式可信链路；4) 在 FaceForensics++基准上进行了系统实验评估，包括对比实验、消融实验，验证了所提方法的有效性和优越性。

2. 相关工作

2.1. 深度伪造被动检测

深度伪造被动检测的方法大致经历了从手工特征到深度学习特征的发展阶段。早期工作主要利用频域分析[12]、纹理特征(如 LBP) [13]-[15]和人脸几何不一致性进行检测，此类方法可解释性强但泛化能力有限。随后，基于 CNN 的端到端检测方法成为主流，Rossler 等[16]提出 FaceForensics++基准并验证了 XceptionNet 的有效性；He 等[17]的 ResNet 架构因其优秀的特征提取能力被广泛用作检测骨干网络[18]。近年来，研究者引入注意力机制[19] [20]和多尺度特征融合[21]以增强对细微伪造痕迹的捕捉；大模型时代下，视觉语言模型(VLM)的多模态推理能力也被初步应用于 Deepfake 检测[22]。然而，已有方法多采用单一检测源或简单加权融合策略，缺乏对多源证据间不确定性和冲突的有效建模。

2.2. 深度伪造主动防御

主动防御旨在从源头阻止或干扰深度伪造生成过程，主要包括数字水印和对抗扰动注入两类方法。数字水印方面，基于 DWT、DCT、SVD 等变换域的盲水印技术[23] [24]可实现内容溯源与版权保护，但单独使用难以阻止伪造过程本身。对抗扰动方面，研究者利用 FGSM、PGD 等对抗攻击方法生成微小扰动以干扰深度伪造模型[25] [26]。裘昊轩[27]研究了针对深度伪造模型的对抗样本生成技术；吴涛[28]探讨了基于对抗样本与区块链的多模态防御体系。然而，现有对抗扰动方法多在整幅图像上均匀添加，未充分利用人脸区域的空间先验信息实现精准分配。本文提出的 GRAP 方法通过梯度引导和显著性图约束，实现了区域自适应的扰动分配策略。

2.3. 多源证据融合

在多源信息融合领域，Dempster-Shafer (D-S)证据理论[29]因其无需先验概率分布且能有效处理不确定性和冲突的特点，已在目标识别、故障诊断等领域获得广泛应用。该理论通过基本概率赋值函数(BPA)描述各证据源的信任分配，利用 Dempster 组合规则实现多源证据的融合。但在深度伪造检测领域，现有融合策略多采用固定权重加权平均，尚未见将 D-S 证据理论应用于多源 Deepfake 检测证据融合的工作。本文首次将 D-S 证据理论引入深度伪造被动检测的多源融合环节，设计了检测概率到 BPA 的映射方法和冲突消解策略，有效提升了融合决策的鲁棒性。

3. 关键技术简介

3.1. 盲水印技术

盲水印是一种在不需要原始图像的情况下就可以提取水印信息的数字水印技术。本系统使用的是基于离散小波变换(DWT)和离散余弦变换(DCT)的盲水印算法。该算法将图像转换为 YUV 色彩空间，然后对 Y 通道进行二级小波分解，在低频子带上进行 DCT 变换，最后通过修改 DCT 系数嵌入水印信息。水印提取时，只需要知道水印的尺寸和密钥就可以恢复出原始水印。相比于传统的图像水印有较强的不可见性、抗攻击性和防篡改性。

$$S'[0]=\left(\left\lfloor\frac{S[0]}{d_1}\right\rfloor+\frac{1}{4}+\frac{1}{2}w_i\right)\times d_1 \tag{1}$$

盲水印算法的核心公式由式(1)表示，其中，$S'[0]$ 为奇异值分解后的最大奇异值，d_1 为量化步长，w_i 为水印比特值。

3.2. 双分支注意力增强检测模型

本文提出双分支注意力增强 ResNet-50 (DBA-ResNet)作为被动检测的核心骨干网络。不同于仅使用通道注意力的传统方案，DBA-ResNet 在 ResNet-50 的第 3、4 阶段并行嵌入空间注意力分支(Spatial Attention Branch, SAB)与通道注意力分支(Channel Attention Branch, CAB)，分别捕获伪造痕迹的空间分布模式和通道响应特征。SAB 通过全局最大池化与平均池化沿通道维度压缩特征图，经 7×7 卷积与 Sigmoid 激活生成空间注意力图；CAB 采用 Squeeze-and-Excitation 结构建模通道间依赖。两分支输出通过可学习权重参数进行自适应融合。

$$\text{Attention}(F)=F\odot\sigma\left(W_2\cdot\text{ReLU}\left(W_1\cdot\text{GAP}(F)\right)\right) \tag{2}$$

通道注意力分支(CAB)的计算过程由式(2)表示，X 为输入特征图，GAP 表示全局平均池化，W_1 和 W_2 为可学习参数，σ 为 Sigmoid 激活函数。

$$F_{out}=\alpha*F_{CAB}+(1-\alpha)*F_{SAB} \tag{3}$$

双分支融合输出由式(3)表示，其中 α 为可学习的融合权重参数，在训练过程中自动调整两个注意力分支的贡献比例。

3.3. 梯度引导区域自适应对抗扰动方法(GRAP)

传统对抗扰动方法(FGSM、PGD 等)在整幅图像上均匀添加扰动，导致背景区域的无效扰动增加视觉失真。本文提出梯度引导的区域自适应对抗扰动方法(GRAP)，其核心思想是利用人脸检测器输出的人脸区域掩码 M 和目标 Deepfake 模型的梯度信息，生成空间自适应的扰动分配权重图 W。具体地，给定原始图像 x，GRAP 首先通过目标模型计算输入梯度 g，计算公式由式(4)表示，然后结合人脸区域掩码 M 构造扰动权重图，如式(5)表示，其中 M_{boundary} 为人脸边界增强掩码，$\lambda_1,\lambda_2,\lambda_3$ 为平衡系数。最终扰动为式(6)所示，其中 ϵ 为扰动预算，扰动预算 ϵ 和迭代次数 N 是决定图像视觉质量与隐蔽性之间平衡的核心超参数，在敏感性实验中，当 ϵ ＜0.03 且 N≤10 时，生成的对抗扰动极其微弱，在经过微信、微博等社交平台的有损压缩后防御成功率急剧下降(低于 40%)；而当 ϵ ＞0.05 时，图像出现人眼可见的光斑与噪点，峰值信噪比(PSNR)降至 28 dB 以下，违背了隐蔽性原则。综合考量后，最终选定 ϵ = 0.03 和 N = 15。GRAP 方法的优势在于：在人脸关键区域集中更强扰动以最大化干扰效果，在背景区域降低扰动以保持视觉质量，同时在人脸边界区域增强扰动以干扰 GAN 的人脸分割与融合过程。此外，系统还叠加频域模式扰动(正弦纹理)，进一步提升对不同生成管线的迁移干扰能力。

$$g=\nabla_x L\left(f(x),y\right) \tag{4}$$

$$W=\lambda_1 M+\lambda_2\text{normalize}\left(|\nabla g|\right)+\lambda_3 M_{\text{boundary}} \tag{5}$$

$$\delta=\epsilon*W*\text{sign}(g) \tag{6}$$

3.4. 基于 D-S 证据理论的多源融合决策

被动检测框架包含传统特征层、深度模型层和视觉大模型层三个独立检测源，如何有效融合其输出

是提升整体检测性能的关键。本文引入 Dempster-Shafer 证据理论替代传统固定权重加权方案。设识别框架 β ={fake, real}，各检测源的检测概率 p_i 通过映射函数转换为基本概率赋值(BPA)：$m_i(\text{fake}) = \beta * p_i$，$m_i(\text{real}) = \beta * (1 - p_i)$，$m_i(\Theta) = 1 - \beta$，其中 β 为信任折扣因子，β 的计算依赖于模型在近期滑动验证集中的历史准确率 Acc_i 以及当前输入的特征置信度熵 H_i，计算公式定义如式(7)表示，其中 α 为权衡系数。多源 BPA 通过 Dempster 组合规则进行融合，公式如式(8)表示，其中归一化因子 K 用于消解冲突证据，K 计算公式如式(9)表示。当冲突系数 K 过大(即证据高度不一致)时，系统自动降低冲突源的信任折扣因子并重新计算。该机制的优势在于：能够自然处理检测结果的不确定性；对单一检测源的误判具有较强鲁棒性；融合结果不仅输出最终判定，还能给出表征不确定性的信任区间，提升决策的可解释性。

$$\beta = \alpha \cdot Acc_i + (1 - \alpha) \cdot \left(1 - \frac{H_i}{H_{\max}}\right) \tag{7}$$

$$m_{1,2}(A) = \frac{1}{K} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{8}$$

$$K = 1 - \sum_{B \cap C = \varnothing} m_1(B) m_2(C) \tag{9}$$

4. 系统设计与实现

本系统的核心技术为盲水印算法、对抗训练和基于深度学习的检测模型，采用 B/S (Browser/Server，浏览器/服务器)架构，如图 1 所示。系统分为表示层、业务逻辑层和数据层三层。表示层使用 Bootstrap 5、Font Awesome 实现响应式用户交互界面；业务逻辑层使用 Flask 框架开发 RESTful API (Application Programming Interface，应用程序编程接口) [30]，实现了主动防御(盲水印、对抗样本)和被动检测两大核心功能模块；数据层使用 SQLite 数据库存储用户信息、检测记录、系统配置等，OSS (Object Storage Service，对象存储服务)存储图片。被动检测模块使用 VLM (Vision-Language Models，视觉语言模型)大模型辅助，加权融合传统检测结果与深度神经网络检测结果。系统还采用了分层日志记录机制，文件和控制台双通道输出，便于防御留痕和系统调试。

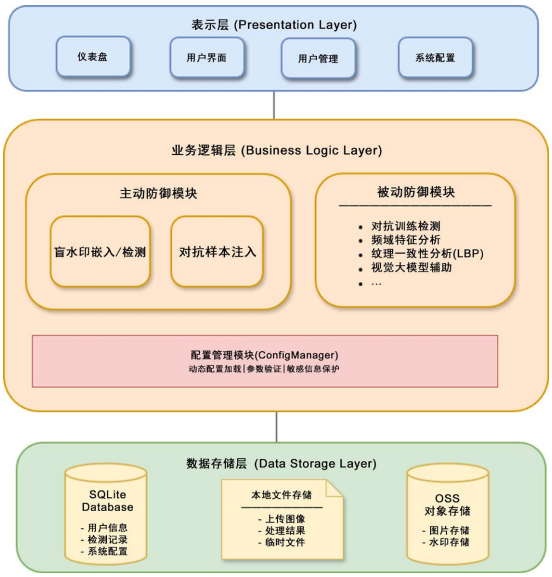

Figure 1. System architecture
图 1. 系统架构图

4.1. 主动防御模块

主动防御模块包括盲水印和对抗样本两个防御子模块。

4.1.1. 盲水印防御模块

盲水印模块是主动防御的重要组成，向原始图像中嵌入隐式的可信水印信息，以达到图像溯源与版权保护的目的。本模块提供/api/embed-watermark 和/api/detect-watermark 两个接口，实现水印隐式嵌入和水印检测提取，流程如图 2 所示。

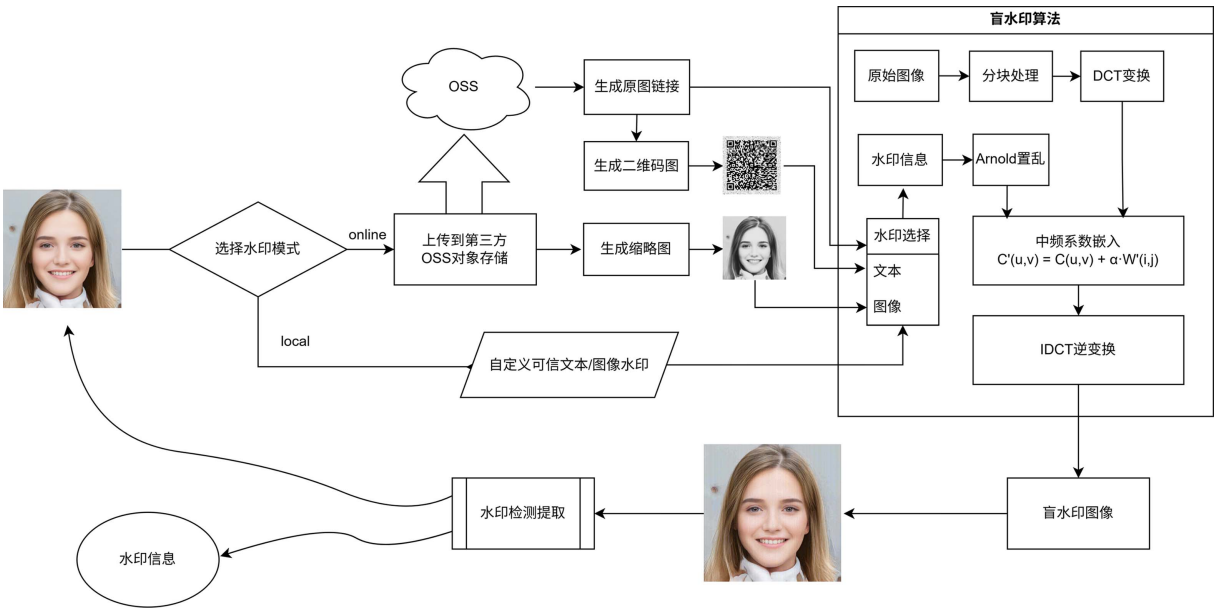

Figure 2. Blind watermark defense module flowchart
图 2. 盲水印防御模块流程图

水印嵌入支持本地模式和在线模式两种工作方式。本地模式支持用户自定义可信文本或图像作为水印信息嵌入原始图像生成水印图，在线模式先上传原始图像至第三方信任云进行 OSS 对象存储，并生成指向原始图像的 url 网络链接，再将该链以文本或生成二维码图像形式作为水印嵌入原始图像生成水印图。所有水印均由用户定义的双重密码加密。底层盲水印算法采用 DWT-DCT-SVD 组合：先对 YUV 通道做 Haar 小波变换得到低频系数，再对 4 × 4 分块系数做 DCT (Discrete Cosine Transform，离散余弦变换)并施加随机置乱(密码种子)，随后通过 SVD (Singular Value Decomposition，奇异值分解)调整奇异值实现比特嵌入。该方案对 JPEG 压缩、轻量噪声与部分几何扰动具有较好的鲁棒性与不可见性平衡。水印提取为嵌入的逆过程，采用 K-means 聚类进行比特判决，提高文本水印的提取准确率。提取过程无需原始图像参与，具有盲提取特性。图 3 为前端用户界面水印嵌入和检测演示。

4.1.2. GRAP 对抗样本防御模块

盲水印防御往往需要搭配其检测功能、构建“嵌入–检测”可信链方能达到良好效果。除此之外，系统还实现了基于对抗样本的主动防御模块，可以最大程度从根本上对抗 Deepfake。该模块通过在原始图像中添加人眼不可见的微小扰动，干扰深度伪造模型的特征提取和生成过程，从而从源头阻止图像被用于深度伪造。系统提供/api/adversarial-defense 接口，将用户上传图像转换为“保护图像”，其视觉效果基本可用，但对深伪模型的人脸特征提取与重演过程造成干扰。图 4 为前端用户交互页面。

(a) (b)

Figure 3. Blind watermark embedding/detection: (a) Example of online text watermark embedding; (b) Example of online QR code watermark extraction
图 3. 盲水印嵌入/提取：(a) 在线文本水印嵌入示例；(b) 在线二维码水印提取示例

Figure 4. Adversarial sample defense
图 4. 对抗样本防御

为保证多样性、提高灵活度，样本可以设置 FGSM、PGD、DeepFool、C&W 四种策略以及轻度、中度、重度三档扰动强度，同时引入人脸区域掩码：优先在人脸区域与边界区域增强扰动，在不显著破坏背景的前提下提升防御针对性。另外，系统还可以叠加频域模式扰动(正弦纹理)与边界扰动(基于掩码梯度的噪声增强)，提高对不同生成管线的迁移干扰能力。从图 5 对比可以看出，原图经过 Deepfake 伪造模型可轻松被篡改，而通过对抗样本注入后的图像与原图并无二致，但经过 Deepfake 模型伪造后则会出现显而易见的干扰，成功抵御了 Deepfake 攻击。

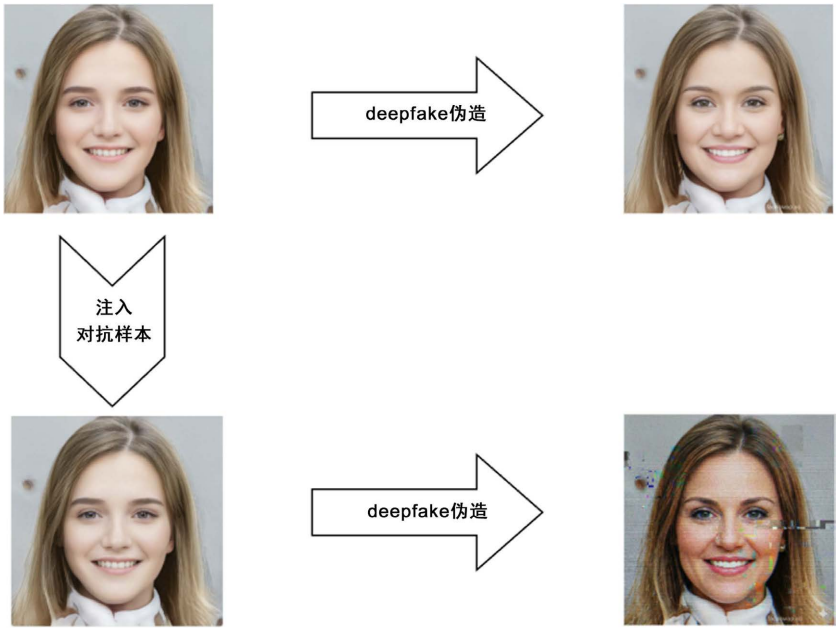

Figure 5. Comparison of adversarial example defenses
图 5. 对抗样本防御对比

4.2. MSEF 被动检测模块

被动检测模块是系统识别深度伪造图像的核心组件，采用“多源证据融合”的设计思想，将传统图像处理方法、深度神经网络与视觉大模型三类检测能力有机整合，形成层次化、可配置、可解释的检测链路(图 6)。

模块整体架构采用三层并行检测与一层融合的设计，输入图像经过预处理后，并行送入以下三个独立的检测分支：

- 传统特征检测层

重点关注图像的物理与统计规律，捕捉深度伪造生成过程中残留的伪影与不自然痕迹。该层包含频域异常分析、纹理一致性检测与人脸对称性分析三个子模块，分别从频谱能量分布、局部二值模式(LBP)纹理特征及面部几何结构三个维度提取特征，输出对应的异常概率 $P_{\text{frequencyp}}$ (频域异常概率)、P_{texture} (纹理异常概率)、P_{face} (特征异常概率)。

- 深度模型检测层

基于 DBA-ResNet 骨干网络的增强型检测器(Enhanced Detector)。为加强对细微伪造痕迹的捕捉能力，该网络使用了空间与通道双分支注意力(Channel Attention)机制，并优化了多层分类头结构。在训练阶段，模型使用 FaceForensics++数据集进行监督学习，并引入对抗训练(Adversarial Training)增强对扰动攻击的鲁棒性。该层负责提取图像的高维隐式特征，输出深度神经网络判定的伪造概率 P_{deep} 及二值化结果 P_{fake} 。

- 视觉大模型检测层

图像上传至 OSS 云生成链接对象作为图像输入，再使用 openai 库远程调用自定义视觉模型 API 接口，设计了控制值域输出的结构化提示词，利用视觉大模型的通用多模态理解能力，从光照一致性、表情自然度及场景逻辑等高层语义维度对图像进行真伪研判。输出语义级的伪造概率 P_{vlm} 。

各检测层独立输出检测概率与置信度，由基于 D-S 证据理论的多源融合层进行自适应聚合。各检测源的概率输出经 BPA 映射后，通过 Dempster 组合规则逐步融合，自动处理证据间的冲突与不确定性，

最终输出结构化检测结果(包含 is_deepfake、confidence、uncertainty_interval、evidence_details 及时间戳等字段)。该架构的核心优势在于：各检测源相互独立、互为补充，融合机制能够自适应处理检测结果间的冲突，信任折扣因子可根据各源的历史精度在线调整，既保证了检测精度又具有良好的可解释性和鲁棒性。

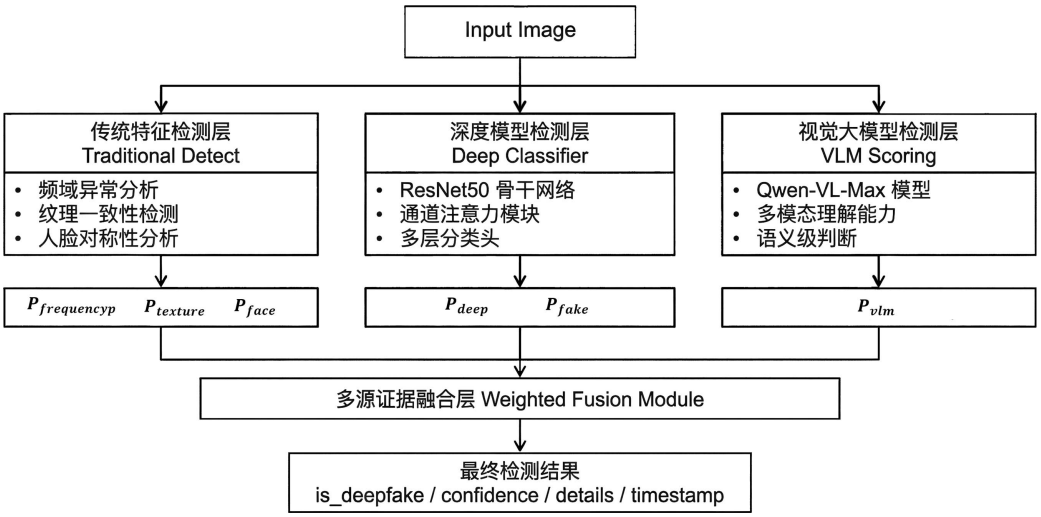

Figure 6. Architecture of passive detection module
图 6. 被动检测模块架构

5. 实验与分析

5.1. 实验设置

本文实验运行环境为 Windows 11，基于 PyTorch 2.2.1 框架，使用 NVIDIA RTX 4090 GPU 进行模型训练。数据集使用广泛用于各种深度伪造检测的 FaceForensics++ (FF++)，其中包含从 Youtube 网站上筛选的大多以新闻播报、独家专访、单人脱口秀等为主题的 1000 段原始视频，仅包含单个人脸，以及分别用 Deepfakes(DF)、Face2Face(F2F)、FaceSwap(FS)和 NeuralTextures(NT)四种篡改方法生成的伪造视频，数据集按压缩质量分为 Raw、C23 和 C40 三个版本，本文采用 C23 版本，对视频帧分解成图片随机抽取进行实验，表 1 为数据集分配情况。输入图像尺寸为 224 × 224。训练时采用 Adam 优化器，初始学习率为 1×10^{-4}，每 20 个 epoch 衰减为原来的 0.1 倍，批量大小为 32，共训练 50 个 epoch。

Table 1. Face Forensics++ dataset allocation
表 1. Face Forensics++数据集分配

FF++数据集		DF	F2F	FS	NT
训练集	Real	72,000	72,000	72,000	72,000
	Fake	72,000	72,000	72,000	72,000
验证集	Real	1400	1400	1400	1400
	Fake	1400	1400	1400	1400
测试集	Real	1400	1400	1400	1400
	Fake	1400	1400	1400	1400

5.2. 对比实验

为验证所提方法的有效性，本文将 MSEF 框架与多种代表性检测方法进行了对比，包括 XceptionNet [16]、MesoNet [31]、F3-Net [32]、基线 ResNet-50 [17]等。表 2 为各方法在 FF++ 数据集上准确率(%)结果。

Table 2. Accuracy results of various methods on the FF++ dataset (%)
表 2. 各方法在 FF++数据集上准确率结果(%)

方法	DF	F2F	FS	NT
ResNet-50 (Baseline)	56.21	63.89	61.25	57.45
XceptionNet	94.30	93.27	87.42	77.89
MesoNet	93.40	92.93	93.51	79.37
F3-Net	91.65	87.03	90.73	60.57
文献[33]	97.80	93.40	88.70	84.20
文献[34]	97.30	94.20	81.80	79.40
MSEF	97.36	97.12	95.58	89.44

由表 2 可知，本文提出的 MSEF 被动检测方法在 F2F、FS 和 NT 三个子数据集上均取得了最优的性能。在 DF 子数据集中，MSEF 达到 97.36%的准确率，较第一名仅差 0.44 个百分点；MSEF 在 F2F 上达到 97.12%，在 FS 上达到 95.58%，在 NT 上达到 89.44%，分别超过次优方法 2.92、2.07 和 5.24 个百分点。这表明 D-S 证据融合策略有效整合了多源异构检测线索，使模型在面对不同伪造方法和数据分布时均具有更强的泛化能力。值得注意的是，在泛化性更具挑战性的 NT 子数据集上，MSEF 的优势更加明显，验证了多源融合对于提升跨域检测鲁棒性的重要作用。

5.3. 消融实验

为验证被动检测模块各层的贡献，本文在 FF++数据集上开展了系统性消融实验。以标准 ResNet-50 为基线，依次叠加空间注意力分支(SAB)、通道注意力分支(CAB)、完整双分支注意力(DBA)、D-S 证据融合和 VLM 语义辅助模块，观察各组件对检测性能的增益。结果如表 3 所示。

Table 3. Ablation experiment results (%)
表 3. 消融实验结果(%)

配置	ResNet-50	SAB	CAB	D-S	VLM	AUC	Acc	F1
Baseline	√	-	-	-	-	95.23	93.18	93.05
-	√	√	-	-	-	96.14	94.27	94.11
-	√	-	√	-	-	96.02	94.09	93.96
DBA	√	√	√	-	-	97.18	95.42	95.30
D-S 证据融合	√	√	√	√	-	98.07	96.53	96.41
MSEF	√	√	√	√	√	98.36	96.81	96.68

分析表 3 发现，单独引入 SAB 或 CAB 均可带来约 0.8%~0.9%的 AUC 提升，说明空间和通道维度的

注意力机制对伪造痕迹的捕捉各有侧重；而 DBA 将两个分支联合使用后，AUC 提升至 97.18%，较基线提高 1.95 个百分点，验证了双分支协同的有效性；D-S 证据融合进一步将 AUC 提升至 98.07%，表明多源异构证据的自适应融合有效缓解了单一检测器的不确定性；VLM 语义辅助模块最终将 AUC 推至 98.36%，证明高层语义信息对底层视觉特征具有良好的互补作用。总体而言，各模块呈正交互补关系，逐步叠加均带来稳定增益。

6. 总结

本文针对深度伪造技术日益增长的安全威胁，提出了一种基于主被动混合策略的深度伪造防御系统——MSEF 框架。在被动检测方面，设计了 DBA-ResNet 双分支注意力增强网络，通过空间注意力与通道注意力的协同作用精准捕捉伪造痕迹；进而提出基于 D-S 证据理论的多源异构证据自适应融合策略，有效整合 DBA-ResNet 视觉特征、频域分析结果及 VLM 语义判断，形成了鲁棒的多源证据融合检测范式。在主动防御方面，提出了 GRAP 梯度引导区域自适应对抗扰动方法，利用面部几何先验知识优化扰动分布，在保持高视觉质量的前提下有效干扰伪造模型；结合鲁棒盲水印技术，实现了溯源追踪与篡改预防的双重保障。在主流 Deepfake 数据集 FF++上的对比实验结果表明 MSEF 被动检测方法综合性能优于现有代表性方法。消融实验验证了 DBA 双分支注意力、D-S 证据融合和 VLM 语义辅助三个核心模块的有效性与互补性。本系统取得了有竞争力的结果，但被动检测系统主要面向已知伪造方法，对未知生成模型的零样本泛化能力有待提升，后续将对其展开相应的深入研究。

基金项目

2025 年江苏省大学生创新训练计划基金项目(xcx2025341, xcx2025350)。

参考文献

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. *NIPS*'14: *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Volume 2, 2672-2680.

[2] 梁俊杰, 韦舰晶, 蒋正锋. 生成对抗网络 GAN 综述[J]. 北京: 计算机科学与探索, 2020, 14(1): 1-17.

[3] 左哲铭. 基于自编码框架的人脸交换方法研究[D]: [硕士学位论文]. 南京: 南京理工大学, 2023.

[4] Mirsky, Y. and Lee, W. (2021) The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, **54**, 1-41. https://doi.org/10.1145/3425780

[5] 王振波, 吴湘玲. 数字时代深度伪造技术研究——机理特征、功能异化及其优化理路[J]. 北京航空航天大学学报社会科学版, 2025, 38(2): 47-55.

[6] 黄晓宝. 基于图像特征提取与融合的深度人脸伪造检测[D]: [硕士学位论文]. 南昌: 南昌大学, 2025.

[7] 姚文达, 李盼池, 赵娅, 等. 人脸深度伪造检测方法研究综述[J]. 中国图象图形学报, 2025, 30(7): 2343-2363.

[8] 瞿左珉, 殷琪林, 盛紫琦, 等. 人脸深度伪造主动防御技术综述[J]. 中国图象图形学报, 2024, 29(2): 318-342.

[9] 刘晓龙, 刘欢, 赵耀, 等. AIGC 伪造内容被动检测与主动防御技术综述[J/OL]. 中国科学: 信息科学, 2025, 55(9): 2250-2288. https://link.cnki.net/urlid/11.5846.TP.20251020.0913.002, 2026-02-13.

[10] 丁峰, 匡仁盛, 周越, 等. 深度伪造及其取证技术综述[J]. 中国图象图形学报, 2024, 29(2): 295-317.

[11] 杨睿, 胡心如, 黄卓超, 等. 深度网络生成式伪造人脸检测方法研究综述[J]. 计算机辅助设计与图形学学报, 2024, 36(10): 1491-1510.

[12] Durall, R., Keuper, M., Pfreundt, F.J., *et al*. (2019) Unmasking Deepfakes with Simple Features. https://arxiv.org/abs/1911.00686

[13] Liu, Z., Qi, X. and Torr, P.H.S. (2020) Global Texture Enhancement for Fake Face Detection in the Wild. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 8057-8066. https://doi.org/10.1109/cvpr42600.2020.00808

[14] 朱新同, 唐云祁, 耿鹏志. 基于特征融合的篡改与深度伪造图像检测算法[J]. 信息网络安全, 2021, 21(8): 70-81.

[15] Abdullah, M.T., Hussein, N. and Ali, M. (2023) Deepfake Detection Improvement for Images Based on a Proposed Method for Local Binary Pattern of the Multiple-Channel Color Space. *International Journal of Intelligent Engineering and Systems*, **16**, 92-104.

[16] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Niessner, M. (2019) Faceforensics++: Learning to Detect Manipulated Facial Images. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 1-11. https://doi.org/10.1109/iccv.2019.00009

[17] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[18] 汤博宇, 焦良葆, 徐逸, 等. 基于改进 ResNet-50 的图像特征提取网络[J]. 计算机测量与控制, 2023, 31(6): 162-167.

[19] 张文祥, 王夏黎, 王欣仪, 等. 一种强化伪造区域关注的深度伪造人脸检测方法[J]. 图学学报, 2025, 46(1): 47-58.

[20] 陈鑫, 高迪, 蒲志明, 等. 基于多模态融合与注意力增强的深度伪造检测方法[J/OL]. 物联网学报, 1-12. https://link.cnki.net/urlid/10.1491.TP.20260210.1057.002, 2026-02-13.

[21] 许楷文, 周翊超, 谷文权, 等. 基于多尺度特征融合重建学习的深度伪造人脸检测算法[J]. 信息网络安全, 2024, 24(8): 1173-1183.

[22] 彭春蕾, 李俊烨, 刘德成, 等. 大模型时代的深度伪造检测[J]. 中国科学: 信息科学, 2026, 56(1): 1-22.

[23] 陈泊睿, 张梅, 李昕蕊, 等. 数字水印技术原理与发展[J]. 中国防伪报道, 2025(10): 98-102.

[24] 高媛. 基于离散小波变换和奇异值分解的数字水印改进算法研究[D]: [硕士学位论文]. 合肥: 安徽建筑大学, 2021.

[25] 陈国凯, 冯辉. 深度学习中对抗样本攻击与防御方法研究[J]. 唐山师范学院学报, 2024, 46(3): 59-66+77.

[26] 刘瑞祺, 李虎, 王东霞, 等. 图像对抗样本防御技术研究综述[J]. 计算机科学与探索, 2023, 17(12): 2827-2839.

[27] 裘昊轩. 针对图像深度伪造模型的对抗样本生成技术研究[D]: [硕士学位论文]. 北京: 中国人民公安大学, 2023.

[28] 吴涛. 生成式 AI 的深度伪造攻击与多模态防御体系研究——基于对抗样本与区块链可追溯验证[J]. 中国信息界, 2025(12): 154-156.

[29] Shafer, G. (1976) A Mathematical Theory of Evidence. Princeton University Press.

[30] 孙业超. 基于 RESTful API 的前后端分离项目接口测试方法研究[J]. 软件, 2025, 46(9): 116-118.

[31] 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, 11-13 December 2018, 1-7. https://doi.org/10.1109/wifs.2018.8630761

[32] 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., *et al*., Eds., *Computer Vision—ECCV* 2020, Springer International Publishing, 86-103. https://doi.org/10.1007/978-3-030-58610-2_6

[33] Jin, X., Wu, N., Jiang, Q., Kou, Y., Duan, H., Wang, P., *et al*. (2024) A Dual Descriptor Combined with Frequency Domain Reconstruction Learning for Face Forgery Detection in Deepfake Videos. *Forensic Science International*: *Digital Investigation*, **49**, Article ID: 301747. https://doi.org/10.1016/j.fsidi.2024.301747

[34] Lin, K., Han, W., Li, S., Gu, Z., Zhao, H., Ren, J., *et al*. (2022) IR-Capsule: Two-Stream Network for Face Forgery Detection. *Cognitive Computation*, **15**, 13-22. https://doi.org/10.1007/s12559-022-10008-4