通过多特征频率感知学习的鲁棒 AI 合成图像检测

Hongfei Cai[1], Chi Liu[1]✉, Sheng Shen[2], Youyang Qu[3,4], and Peng Gui[5]

[1] Faculty of Data Science, City University of Macau, Macao SAR, China
[2] Design and Creative Technology Vertical, Torrens University Australia, NSW, Australia
[3] Shandong Provincial Key Laboratory of Computer Networks, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[4] School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China
✉ Corresponding author: chiliu@cityu.edu.mo

摘要 生成式 AI（GenAI）技术的迅速发展加剧了人们对滥用 AI 生成图像的担忧。为了解决这一问题，出现了强大的检测方法，尤其是在目标 GenAI 模型超出分布范围或生成的图像在传输过程中受到扰动等具有挑战性的条件下尤为突出。本文介绍了一个多特征融合框架，旨在通过引入三个互补组件——噪声相关性分析、图像梯度信息和预训练视觉编码器知识，并采用跨源注意力机制来增强空间取证特征表示。此外，为了识别合成图像中的频谱异常，我们提出了一种频率感知架构，该架构使用了频率自适应扩张卷积，能够在保持低计算复杂度的同时联合建模空间和频谱特征。我们的框架在包括文本到图像扩散模型、自回归方法和后期处理的深度伪造方法在内的十四种多样化的 GenAI 系统中表现出色。值得注意的是，在跨模型检测任务中，与现有的最先进的技术相比，它实现了显著更高的平均准确率。此外，提出的方法对各种类型的真实世界图像噪声扰动（如压缩和模糊）显示出抗干扰能力。广泛的消融研究进一步证实了融合多模态取证特征与频率感知学习的协同效益，突显了我们方法的有效性。

Keywords: AI 合成图像。鲁棒检测。特征融合

1 介绍

生成人工智能（GenAI）模型，如生成对抗网络（GANs）和扩散模型的最新进展使对手能够以低成本制造逼真的图像。这给数字完整性与真实性带来了重大威胁。作为回应，检测 AI 合成图像已成为保护 AI 生成内容的初始防线中越来越关键的一环。

在人工智能合成图像检测中仍存在两个持续的挑战：泛化能力和检测器的鲁棒性。随着生成假图像的 GenAI 模型不断发展，针对特定 GenAI 模型训练的检测器应能够泛化以识别先前未见过的 GenAI 模型。此外，鉴于在线分布的人工智能生成图像可能会受到各种传输噪声（如压缩和模糊）的影响，训练有素的检测器必须对典型的图像扰动保持鲁棒性。

最近，提出了各种人工智能合成图像检测器，利用了来自图像域的取证特征，如纹理细节 [2]，噪声关系 [25]，和梯度 [26]，以及频率域中的频谱分布 [8]，傅里叶振幅 [28]，和 DCT 系数 [8]。此外，一些方法采用特征融合来整合这些多样化的特征来源 [9]。然而，在解决跨模型泛化和噪声鲁棒性问题上仍然存在差距。依赖单一特征源的方法容易受到目标 GenAI 模型变化的影响，特别是在测试 GenAI 模型架构发生变化时，例如从 GANs 变为扩散模型。此外，某些特征可能特别敏感于特定的噪声扰动；例如，在线图像压缩将显著损害原始频域特征表示 [8]。多特征融合策略为解决单特征表示的局限性提供了有希望的方法途径；然而，仍然需要更复杂且有效的融合机制，这值得进一步研究。

为了解决这些挑战，我们提出了一种新颖的多特征融合和频率感知学习框架，用于通用且稳健的人工合成图像检测。我们的方法首先通过集成噪声关系特征、图像梯度特征以及来自预训练大型视觉编码器的知识，并结合跨源注意力模块，创建了一个强大的取证多特征表示。噪声和梯度特征提供了可靠的取证线索 [26, 25]，而表征自然图像分布的预训练知识通过锐化自然图像与人工图像之间的决策边界来增强特征表示 [22]。集成后的特征随后被输入到频率感知学习主干网络中。频率感知学习主干涉及将残差学习与频率自适应扩张卷积（FADC）结合 [3]。这种设计提高了检测器同时捕获之前融合特征中的空间信息和频率信息的能力，特别关注局部频率动态变化。同时，FADC 有助于减少网络复杂度并提高计算效率以获得轻量级检测器。

为了验证所提出的多特征频率感知学习框架的有效性，我们进行了广泛的实验，针对十四生成式 AI 模型的检测，涵盖了类似于 GAN 的模型、扩散式的文本到图像模型、自回归模型、低级和感知图像处理模型以及后期渲染的深度伪造。我们还将我们的方法与各种方法进行比较，包括广泛使用的分类器如 ResNet 和 ViT，以及一些最近的状态-of-the-art 基线。此外，还提供了消融研究以展示多特征融合的合理性。结果证实，我们的方法在检测 AI 合成图像方面实现了卓越的准确性、泛化能力和鲁棒性。

总结而言，本文的贡献包括：

- 我们提出了一种新的 AI 合成图像检测框架，该框架显著提高了检测的跨模型泛化能力和噪声鲁棒性。
- 我们设计了一种多特征融合机制，使模型能够自适应地结合噪声关系、图像梯度以及预训练大型视觉编码器的知识，以实现稳健的取证特征表示。
- 我们引入了一种频率感知的学习骨干网络，该网络通过轻量级设计以较低的计算成本有效整合全局空间和局部频率信息。
- 我们广泛的实验评估表明，我们的框架在准确性、泛化能力和鲁棒性方面表现出色，超越了多种基线方法，在各种测试设置中均取得了优异的结果。

2 相关工作

基于图像的检测方法 专注于分析图像中的像素级和区域特征以识别潜在的伪造。这些特征包括纹理细节 [2]、色彩分布 [10]、饱和度 [18]、边缘信息 [21] 和其他视觉线索。例如，一些研究探讨了使用梯度信息来检测图像纹理和边缘中的不一致性，这些不一致性是 AI 生成内容的特征 [21]。这些方法特别有效于识别合成过程中引入了细微但可检测的空间域异常的伪造品。然而，它们可能对图像质量和分辨率敏感，并且可能难以处理那些非常接近真实图像的高质量伪造品。例如，最近的研究表明，可以通过追踪移除攻击有效消除空间异常，这突显了仅依赖空间特征进行鲁棒检测的局限性 [14]。

基于频率的检测方法 将图像从空间域转换到频率域，利用高频伪影等特征来识别合成内容 [8]。这些方法利用了 AI 生成的图像通常表现出独特的频率模式，如周期结构或高频成分中的异常，而这些在自然图像中通常是不存在的。例如，使用傅里叶变换或小波变换可以提取频率特征，从而揭示合成元素的存在 [8]。进一步探索了 GenAI 模型的任务特定取证的频率域指纹识别 [16]。基于频率的方法对于检测留下独特频率异常的 GenAI 模型是有用的；然而，它们可能高度易受图像噪声或其他频率域干扰的影响 [29]。

特征融合检测方法 结合来自不同层次的各种图像特征，如视觉特征、频率特征或多模态数据，以提高检测的准确性和鲁棒性 [9, 11]。通过整合多源信息，这些方法旨在捕捉图像的空间和频域特性，从而提升检测性能。例如，一些研究提出了融合全局和局部特征的框架，利用不同特征集的互补性来实现对各种伪造类型和数据集更好的泛化 [9]。此外，特征融合方法通常会融

入先进的技术如注意力机制，动态权衡不同特征的重要性，进一步优化检测过程 [11]。这些方法一般对于伪造技术和图像质量的变化更加鲁棒，在遇到各种伪造场景的实际应用中更为适用。提出了一种新颖的多视角完整表示法，用于增强 GAN 生成图像的检测稳健性，有效整合了多尺度特征和跨视角信息 [15]。

3 方法

我们的框架用于检测 AI 合成图像，将四个互补的特征表示整合在一个频率感知架构中，同时处理空间和频谱取证模式。如图 1所示，该设计动机源于三个关键观察：(i) 现代生成模型尽管具有视觉逼真度但仍表现出语义不一致，需要进行全局语义分析；(ii) 篡改伪迹表现为局部空间不连续性；(iii) 合成图像通常在子频带中包含频率域异常。因此，该流水线采用三个并行处理分支，然后进行频率域细化。CLIP-ViT 模块利用预训练的视觉语言嵌入来捕捉语义一致性，这得益于它在大规模数据集上证明了跨模式对齐能力。同时，基于变换的空间梯度分析器通过 Sobel 算子检测局部像素级不一致，针对生成图像边界中的常见伪迹。第三个分支通过引导滤波计算噪声图案残差（NPR），以隔离纹理区域的结构异常。

这些异构特征被拼接并通过频率选择模块进行处理，该模块应用离散小波变换（DWT），将信号分解为近似部分（低频）和细节部分（高频）。这种设计明确地建模了频域-空间交互作用，因为合成图像在特定子带中经常表现出异常的能量分布。随后的全局残差块采用身份映射来放大判别模式并减轻梯度消失问题。空间注意力机制根据局部伪影严重程度动态重新加权特征图，将计算集中在可疑区域上。网络架构利用堆叠的 3Œ3 卷积层和 ReLU 激活函数及批量归一化，通过层次处理逐步抽象出特征，在最终使用全连接层进行二元分类之前。这种多阶段设计允许联合建模互补的取证线索，同时通过模块化组件保持参数效率。

3.1 相邻像素关系和梯度特征

邻像素关系(NPR)： 受生成网络 [25] 中固有的上采样模式的启发，我们通过基于网格的差分算子显式建模局部像素相关性。给定输入图像 $I \in \mathbb{R}^{H\times W\times 3}$，我们首先将其划分为 $W \times H$ 个规则网格 $\{v_c^I\}_{c=1}^{W\times H}$，其中每个网格包含 $l \times l$ 个像素（对于常见的 2× 上采样为 $l = 2$）。通过每个网格内的全面成对差计算 NPR 特征张量 $\hat{V}^I \in \mathbb{R}^{W\times H\times (l^2-1)}$：

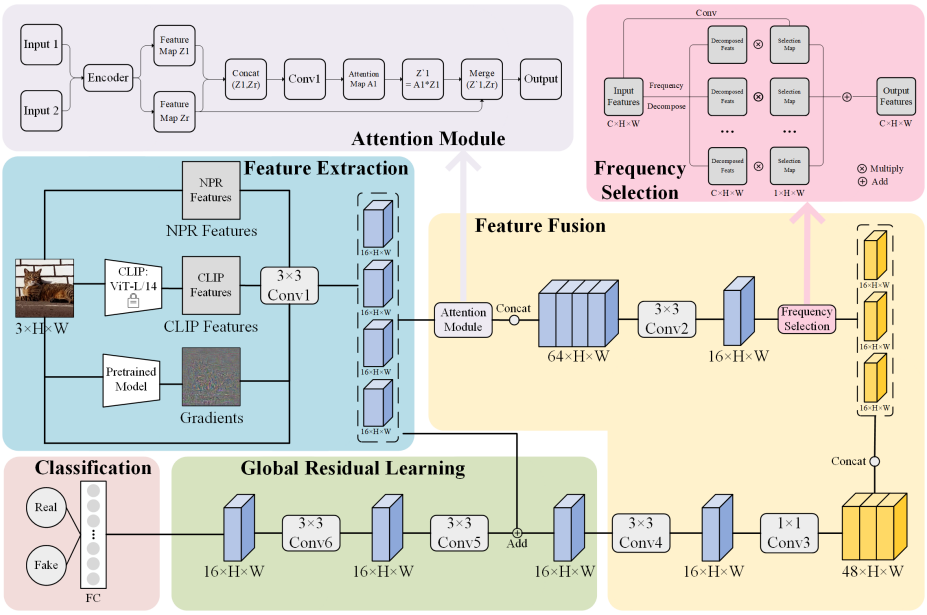

图 1: 所提出的多特征频率感知学习模型的架构。多分支网络集成了 CLIP 语义特征、变换梯度和 NPR（噪声模式残差）特征。通过离散小波变换进行频率分解，分离低频和高频成分，随后进行残差增强特征细化和基于注意力的特征加权。最终分类是通过分层卷积块和全连接层实现的。

$$\hat{v}_c^I = \left\{ w_i - w_j \mid \forall w_i \in v_c^I,\ 1 \le j \le l^2 \right\} \quad \forall c \in \{1, \ldots, W \times H\} \tag{1}$$

其中 I 表示具有高度 H，宽度 W 和 3 个颜色通道的输入图像。网格 $\{v_c^I\}$ 从图像中划分出来，每个包含 $l \times l$ 像素。所得的 NPR 特征 $\hat{V}^I$ 通过计算每个网格 v_c^I 内的成对差值来捕捉局部像素相关性。网格内的像素值由 w_i 和 w_j 表示，c 是网格的索引。此操作放大了 GAN 中转置卷积产生的特征网格伪影，特别是在假周期纹理如头发丝和织物编织中尤为明显。选择 $l = 2$ 对应于现代生成器中最常见的上采样因子。

梯度特征：为了捕捉互补的伪影特征，我们使用一个固定的 CNN 主干网络 M（在 ImageNet 上预训练的 ResNet-50）计算梯度图。对于每个输入图像 $I_i \in \mathbb{R}^{H \times W \times 3}$，我们通过指导反向传播计算梯度响应 $G \in \mathbb{R}^{H \times W \times 3}$：

$$G = \frac{\partial \sum_{k=1}^{K} M_k(I_i)}{\partial I_i} \tag{2}$$

其中 M 表示固定的 CNN 主干（ResNet-50），I_i 是输入图像，G 是梯度响应图。M_k 代表最后一个卷积层的第 k 个通道输出，而 K 是该层中的通道数量。梯度特征突出了扩散模型输出中的高频伪影，同时通过保持 CNN 参数不变来防止过拟合训练数据统计信息。这些特征特别适用于检测合成阴影和反射的边界不连续性。

3.2 融入预训练的语义先验

我们将来自 CLIP-ViT-L/14 的语义先验 [22] 整合以区分真实的自然图像统计信息。给定一个图像 x，我们从投影头之前的最终变换器层中提取 CLIP 特征向量 $\phi_x \in \mathbb{R}^{768}$，保留了语义和纹理信息。

一个多头交叉注意力机制动态融合了 CLIP 的全局语义与局部特征（NPR⊕ 梯度）。设 $F_{\text{local}} = [\hat{V}^I \oplus G] \in \mathbb{R}^{L \times C}$ 表示拼接后的局部特征（$L = W \times H$ 空间位置，C 通道）。融合过程形式化为：

$$\begin{aligned} Q &= W_Q \phi_x \in \mathbb{R}^{d_k} \\ K &= W_K F_{\text{local}} \in \mathbb{R}^{L \times d_k} \\ V &= W_V F_{\text{local}} \in \mathbb{R}^{L \times d_v} \\ \text{Attention}(Q, K, V) &= \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \in \mathbb{R}^{d_v} \end{aligned} \tag{3}$$

其中 W_Q, W_K, W_V 是可学习的投影。softmax 温度 $\sqrt{d_k}$ 在训练过程中稳定梯度流动。这种注意力机制使自适应特征重校准成为可能，CLIP 的语义向量（查询）通过与键的兼容性分数选择性地放大判别局部特征（值）。

3.3 频率感知残差学习

在我们的频率感知主干中，每个频率自适应扩张卷积（FADC）块旨在通过多路径架构共同建模空间和频谱信息，该架构包括频率选择以平衡高低频分量。详细公式如下：

高频路径

$$Y_{\text{FADC}}(p) = \sum_{k=1}^{K^2} \left(w_k^{\text{low}} + w_k^{\text{high}} \cdot \lambda_h(p) \right) \cdot X(p + \Delta p_k \times \hat{D}(p)) \tag{4}$$

其中 $Y_{\text{FADC}}(p)$ 表示来自高频路径的空间位置 p 的输出特征，其中 K 是卷积核大小，w_k^{low} 和 w_k^{high} 分别是分解后的低频和高频成分的卷积权重。$\lambda_h(p)$ 是一个用于高频分量的动态权重因子，$\hat{D}(p)$ 是自适应膨胀率。

低频路径

$$Y_{\text{skip}}(p) = X(p) \tag{5}$$

此路径通过跳过连接直接保留了输入特征图 X 的低频内容。

频率适应模块

$$\hat{D}(p) = \text{ReLU}(f_\theta(X(p))) \times D_{\text{base}} \tag{6}$$

其中，$\hat{D}(p)$ 是根据输入特征的局部频率内容动态调整的自适应扩张率。f_θ 是一个实现为深度可分离卷积层的轻量级频率预测器，而 D_{base} 是预定义的基础扩张率。

频率选择模块

$$X_b = \mathcal{F}^{-1}(M_b \cdot \mathcal{F}(X)) \tag{7}$$

$$\hat{X}(p) = \sum_{b=0}^{B-1} A_b(p) \cdot X_b(p) \tag{8}$$

其中 $\mathcal{F}$ 和 $\mathcal{F}^{-1}$ 分别表示傅里叶变换和逆傅里叶变换。M_b 是一个二进制掩码，用于提取第 b 个频率带，而 $A_b(p)$ 是第 b 个频率带的空间变权重图。$\hat{X}(p)$ 是频率平衡特征图。

输出积分

$$Y_{\text{out}}(p) = \text{ReLU}(Y_{\text{FADC}}(p) + Y_{\text{skip}}(p) + \hat{X}(p)) \tag{9}$$

最终输出 $Y_{\text{out}}(p)$ 集成了来自 FADC 路径的高频特征、来自跳跃连接的低频特征以及来自频率选择模块的频率平衡特征。

3.4 优化目标

网络使用类别平衡焦损函数进行端到端训练：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha y_i(1-p_i)^{\gamma}\log p_i + (1-\alpha)(1-y_i)p_i^{\gamma}\log(1-p_i)] \qquad (10)$$

其中 $p_i = \sigma(\text{logit}_i)$ 是预测概率，$\alpha = \frac{N_{\text{fake}}}{N_{\text{real}}+N_{\text{fake}}}$ 平衡类别频率，而 $\gamma = 2$ 将训练重点放在难例上。温度参数 γ 使损失景观平滑以提高收敛稳定性。

4 实验

4.1 目标生成模型

由于制造假图像的新方法不断出现，使用一个生成模型的图像进行训练是标准做法，并测试该模型检测来自其他未见过的模型的假图像的能力也是标准做法。我们遵循 [27] 中概述的协议，使用 ProGAN[12] 的真实和假图像进行训练。在评估期间，我们考虑一组多样化的生成模型。最初，我们在参考的模型上评估性能：[27]：ProGAN、StyleGAN[13]、BigGAN[1]、CycleGAN[30]、StarGAN[4] 和 GauGAN[23]。每个生成模型为我们的分析提供了真实和虚假图像的不同集合。此外，我们将评估扩展到包括在 ImageNet 数据集上训练的引导扩散模型 [7] 以及近期的文本到图像生成模型：潜在扩散模型（LDM）[24]、Glide[20]，以及自回归模型 DALL-E[6]。

4.2 基线检测器

我们将我们的方法与几个最先进的基线进行了比较：(i) 使用二元交叉熵损失在 ProGAN 生成的真实和假图像上训练的分类网络，具体细节见 [27]。该网络利用了 ImageNet 预训练的 ResNet-50。我们还考虑了一种变体方法，其中主干网络更改为 CLIP-ViT 以与我们的特征空间方法保持一致。(ii) 提出于 [2] 的一种基于补丁级别的分类方法，该方法截断了 ResNet 或 Xception 网络，以便在将图像分类为真实或假时专注于较小的感受野。(iii) 在真实和假图像的共现矩阵上训练的分类网络，在图像隐写分析和取证中已被证明是有效的 [5]。(iv) 基于真实和假图像频率谱的分类网络，捕捉 GAN 生成图像中存在的伪影 [28]。(v) 由 [22] 提出的方法，该方法利用 CLIP-ViT 的特征空间进行最近邻分类（NN）或线性分类（LC）。我们专注于比较 NN 变体。

4.3 评估指标和设置

我们的实验配置遵循了既定的多媒体取证协议 [2, 19, 27, 28]，通过分类准确率（ACC）和平均精度（AP）分数来评估检测性能。为了确保公平比较，我们使用了一个包含所有目标生成模型样本的保留验证集来校准分类阈值。

训练范式采用带有动量 0.9 的随机梯度下降，初始学习率为 1×10^{-4}，批量大小为 32，处理 3 通道的 256×256RGB 图像。我们对前 500 次迭代实施线性预热，随后进行余弦学习率衰减。模型训练了 20 个周期，使用交叉熵损失函数，并使用随机种子 1 进行确定性权重初始化。该架构利用 Adam 优化（$\beta_1 = 0.9$，$\beta_2 = 0.999$）并带有权重衰减 5×10^{-4}，在卷积层之后加入了批量归一化和 dropout（$p = 0.2$）。所有实验在 PyTorch 2.0 上运行，使用混合精度训练，并在 NVIDIA RTX 4090 GPU 上进行评估。

4.4 检测的泛化

提出的多特征频率感知学习框架展示了出色的泛化能力，特别是在分布外（OOD）生成模型上，包括 GAN 和扩散模型。如表 1和表 2所示，我们的方法在多个未见过的 GAN 和扩散模型上实现了高精度，显著优于几个最先进的基线方法。这表明该框架即使是在单个生成模型（ProGAN）训练并在多种不同的未知架构上测试时，也非常有效地检测 AI 合成图像。

这些 OOD 扩散模型的优越性能值得注意，因为这些模型以其高质量和逼真的视觉效果而闻名，这往往对现有的检测方法构成挑战。我们的框架能够跨如此多样且先进的生成模型进行泛化，凸显了其鲁棒性和适应性。我们方法的有效性归功于多模态特征的集成以及频率感知学习主干。这些组件使模型能够捕捉到 AI 生成图像特有的空间和光谱异常，从而增强了自然图像与合成图像之间的决策边界。表 3中的结果进一步支持了这一结论，展示了每个特征分支在实现稳健检测中的重要性。此外，图 2中的 t-SNE 可视化说明，在引入频率域模块后，真实和伪造图像间的分离得到了改进，证实了该框架增强特征区分能力的能力。

4.5 抗后期处理操作性

为了评估我们的分类器对抗潜在规避策略的鲁棒性，我们评估了它们在常见图像处理操作（如 JPEG 压缩和高斯模糊）下的性能。我们从 [27] 中选择基线，因为它是一种广泛采用的增强 AI 生成内容检测中图像扰动鲁棒性

表 1: 各种 GAN 模型检测的平均准确率（ACC）。检测器是使用真实图像和 ProGAN 图像进行训练的。

Detection Method	ProGAN	CycleGAN	BigGAN	StyleGAN	GauGAN	StarGAN	Total
CNN Detection [27]	98.94	78.80	60.62	60.56	66.82	62.31	71.34
Patch Classifier [2]	94.38	67.38	64.62	82.26	57.19	80.29	74.35
Co-occurrence [19]	97.70	63.15	53.75	**92.50**	51.10	54.70	68.82
Freq-spec [28]	44.90	**99.90**	50.50	49.90	50.30	**99.70**	65.87
UniFD [22]	99.54	93.49	88.63	80.75	**97.11**	98.97	**93.08**
Ours	**99.64**	90.11	**92.55**	87.72	94.29	93.34	92.94

表 2: 检测由各种扩散模型生成的图像的平均准确率（ACC）。检测器使用真实图像和 ProGAN 图像进行训练。

Detection Method	LDM			Glide			Guided	DALL-E	Total
	200 steps	200w/CFG	100 steps	100 27	50 27	100 10			
CNN Detection [27]	50.74	51.04	50.76	52.15	53.07	52.06	50.66	53.18	51.71
Patch Classifier [2]	79.09	76.17	79.36	67.06	68.55	68.04	65.14	69.44	71.61
Co-occurrence [19]	70.70	70.55	71.00	70.25	69.60	69.90	60.50	67.55	68.76
Freq-spec [28]	50.40	50.40	50.30	51.70	51.40	50.40	50.90	50.00	50.69
UniFD [22]	91.29	72.02	91.29	89.05	90.67	90.08	71.06	81.47	84.62
Ours	**94.40**	**80.60**	**94.30**	**91.70**	**91.10**	**91.90**	**95.10**	**86.90**	**90.75**

表 3: CLIP 特征、梯度和 NPR 特征对检测的影响通过展示移除特征后的性能变化来说明。“−”表示已排除特征分支。“ALL”表示完整模型的性能。

Model	$-f_{Clip}$	$-f_{Grad}$	$-f_{NPR}$	**ALL**
GAN models	74.23	53.45	62.17	**90.56**
Diffusion models	91.34	55.67	61.89	**92.78**

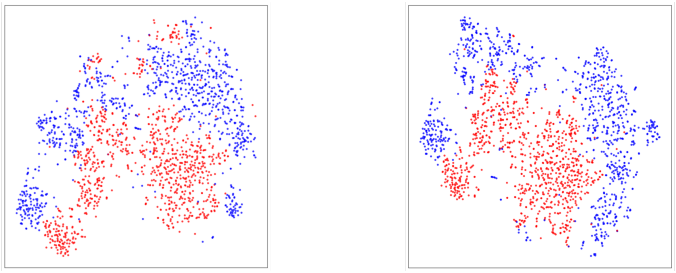

图 2: t-SNE 可视化了 2000 张测试图像，展示了频率域模块在提高真实（蓝色）和伪造（红色）图像二分类准确性方面的有效性。左侧面板表示没有使用频率域模块的特征分布，而右侧面板显示了应用该模块后改进的分离效果。

的策略，并且提供了一个成熟的基准。图 3展示了在不同水平的 JPEG 压缩和高斯模糊下我们的方法与基线方法的比较结果。相比于基线，我们的方法表现出更高的 AP 分数，特别是在分布外检测（即检测器是在 ProGAN 图像上训练但在 Diffusion 图像上测试）的情况下。这突显了我们方法在现实世界挑战场景中的优越性，在这些场景中生成模型未知且图像可能在传播过程中受到各种扰动。另一个不寻常的观察是，增加扰动强度与检测扩散模型生成图像时更高的 AP 分数相关联。一个可能的原因是由于扩散模型 [17] 固有的去噪过程，扩散图像对外部扰动的敏感度低于真实图像；这仍需进一步探索。

4.6 讨论

实验结果强调了我们提出的假图像检测方法的有效性。该方法不仅在各种生成模型中表现出强大的性能，还显示出对抗后期处理操作的韧性。这些能力使其成为打击日益蔓延的假图像传播的强大工具。尽管图表中出现了意料之外的趋势，即随着图像质量下降和 Sigma 增加，我们的模型性能似乎有所提高，我们认为这可能是归因于扩散模型生成过程中固有的噪声添加过程。这种噪声可能使在较低质量下产生的图像与真实图像更加可区分，从而有可能增强我们检测模型识别它们为假的能力。需要进一步调查该模型在这种条件下的行为以完全理解和解释这些结果。这可能涉及检查扩散模型引入的特定噪声特征及其如何与检测模型的特性相互作用。

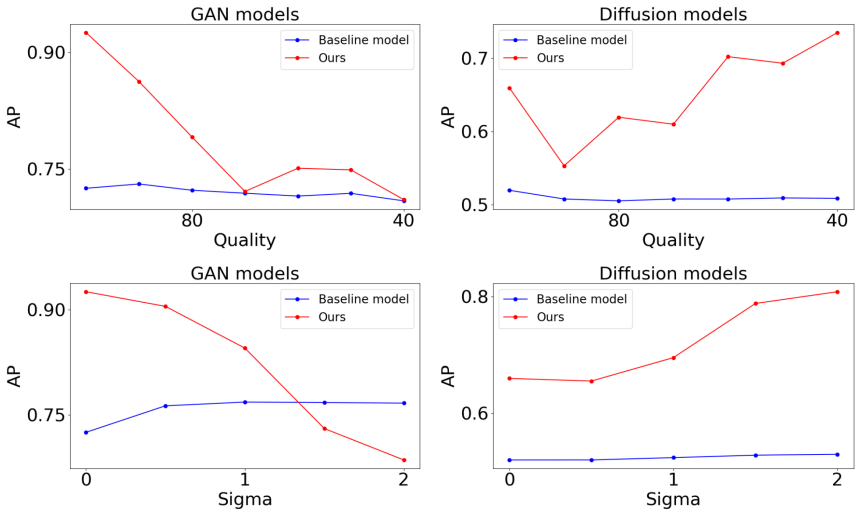

图 3: 不同检测方法对各种图像压缩和噪声的鲁棒性。基线模型 [27] 和我们的模型在两种生成模型（GAN 和 扩散）下，不同质量的 JPEG 压缩（顶部行）和不同级别的高斯噪声（底部行）下的平均 AP 分数进行了比较。

5 结论

总之，人工智能合成图像生成技术的快速发展对数字内容的真实性和可靠性构成了重大挑战。本研究提出了一种基于频率感知多特征融合的稳健的人工智能合成图像检测框架，旨在应对这些复杂伪造技术所带来的威胁。通过将空间、梯度、频域和 CLIP 特征集成到深度学习架构中，我们的模型利用多样化的视觉表示来提高检测精度。采用注意力机制能够动态学习特征的重要性，优化检测性能。在综合数据集上的大量实验表明，本方法具有卓越的准确性、泛化能力和对各种攻击场景的鲁棒性。随着人工智能合成图像生成技术的发展，我们框架的适应性和鲁棒性使其成为对抗虚假图像传播的一件强大工具。这项研究贡献于当前最先进的 AI 合成图像检测领域。它为未来研发更通用和可靠的检测解决方案奠定了坚实的基础。

参考文献

1. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. CoRR **abs/1809.11096** (2018)

2. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 103–120. Springer International Publishing, Cham (2020)
3. Chen, L., Gu, L., Zheng, D., Fu, Y.: Frequency-adaptive dilated convolution for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3414–3425 (June 2024)
4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Cozzolino, D., Poggi, G., Verdoliva, L.: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. p. 159 – 164. IH&MMSec '17, Association for Computing Machinery, New York, NY, USA (2017)
6. Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., Melas, L., Ghosh, R.: Dall · e mini (7 2021)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
8. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3247–3258. PMLR (13–18 Jul 2020)
9. Guo, Z., Jia, Z., Wang, L., Wang, D., Yang, G., Kasabov, N.: Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. IEEE Transactions on Information Forensics and Security **19**, 401–413 (2024)
10. He, P., Li, H., Wang, H.: Detection of fake images via the ensemble of deep representations from multi color spaces. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2299–2303 (2019)
11. Ju, Y., Jia, S., Cai, J., Guan, H., Lyu, S.: Glff: Global and local feature fusion for ai-synthesized image detection. IEEE Transactions on Multimedia **26**, 4073–4085 (2024)

12. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. CoRR **abs/1710.10196** (2017)

13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

14. Liu, C., Chen, H., Zhu, T., Zhang, J., Zhou, W.: Making deepfakes more spurious: evading deep face forgery detection via trace removal attack. IEEE Transactions on Dependable and Secure Computing (2022)

15. Liu, C., Zhu, T., Shen, S., Zhou, W.: Towards robust gan-generated image detection: a multi-view completion representation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023). pp. 464–472 (2023)

16. Liu, C., Zhu, T., Zhao, Y., Zhang, J., Zhou, W.: Disentangling different levels of gan fingerprints for task-specific forensics. Computer Standards & Interfaces **89**, 103825 (2024)

17. Luo, Y., Du, J., Yan, K., Ding, S.: Lare$\hat{2}$: Latent reconstruction error based method for diffusion-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17006–17015 (2024)

18. McCloskey, S., Albright, M.: Detecting gan-generated imagery using saturation cues. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 4584–4588 (2019)

19. Nataraj, L., Mohammed, T.M., Manjunath, B.S., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K.: Detecting GAN generated fake images using co-occurrence matrices. CoRR **abs/1903.06836** (2019)

20. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. CoRR **abs/2112.10741** (2021)

21. Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: Deepfake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6111–6121 (2022)

22. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24480–24489 (June 2023)

23. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)

25. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28130–28139 (June 2024)

26. Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12105–12114 (June 2023)

27. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

28. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2019)

29. Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., Yu, P.S.: Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. ACM Computing Surveys **55**(8), 1–39 (2022)

30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)