Computer Science and Application 计算机科学与应用, 2026, 16(11), 28-41
Published Online November 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.1611323

Hans汉斯

基于通道－空间自适应与多残差环注意力机制的人工智能生成图像检测

彭时雨

重庆工商大学数学与统计学院，重庆

收稿日期：2026年10月1日；录用日期：2026年10月29日；发布日期：2026年11月9日

摘　要

针对生成式人工智能快速发展导致的AI生成图像真实性验证难题，以及现有检测方法依赖单模态特征、泛化能力有限的问题，旨在提升对未知生成模型和复杂后处理操作的检测准确率与鲁棒性。首先设计了通道－空间注意力自适应残差模块，通过双重注意力机制和自适应融合策略增强图像的高频噪声、边缘梯度和频域变换关键特征；再引入残差环增强的多头注意力机制，在标准多头注意力计算之后串联残差环，对注意力输出进行迭代式非线性后处理，提高特征交互的表示复杂度；最后进行多层次特征的稳定融合。在GenImage数据集上的实验结果表明，所提方法对七种主流生成模型生成的图像检测准确率达到99.67%，显著优于现有最先进方法。该方法展现了优越的泛化能力和鲁棒性，能够有效应对未知生成模型与复杂后处理操作，为AI生成图像检测提供了高效可靠的技术途径。

关键词

AI生成图像检测，多模态特征融合，自适应注意力机制，通道－空间注意力，残差环注意力

AI-Generated Image Detection Based on Channel-Spatial Adaptation and Multi-Residual Ring Attention Mechanism

Shiyu Peng

School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing

Received: October 1, 2026; accepted: October 29, 2026; published: November 9, 2026

Abstract

To address the challenge of authenticating AI-generated images arising from the rapid development

文章引用: 彭时雨. 基于通道-空间自适应与多残差环注意力机制的人工智能生成图像检测[J]. 计算机科学与应用, 2026, 16(11): 28-41. DOI: 10.12677/csa.2026.1611323

of generative artificial intelligence, and to overcome the limitations of existing detection methods that rely on single-modal features and exhibit limited generalization capability, this study aims to improve detection accuracy and robustness against unknown generative models and complex post-processing operations. We first design a channel-spatial attention adaptive residual module, which enhances key features-including high-frequency noise, edge gradients, and frequency-domain transform characteristics-through a dual-attention mechanism and an adaptive fusion strategy; then we introduce a residual ring-enhanced multi-head attention mechanism that appends a residual ring after the standard multi-head attention computation to perform iterative nonlinear post-processing on the attention output, thereby increasing the representational complexity of feature interactions; and finally we perform stable fusion of multi-level features. Experimental results on the GenImage dataset demonstrate that the proposed method achieves a detection accuracy of 99.67% on images generated by seven mainstream generative models, significantly outperforming existing state-of-the-art methods. The proposed method exhibits superior generalization capability and robustness, effectively handling unknown generative models and complex post-processing operations, thus providing an efficient and reliable technical approach for AI-generated image detection.

Keywords

AI-Generated Image Detection, Multi-Modal Feature Fusion, Adaptive Attention Mechanism, Channel-Spatial Attention, Residual Ring Attention

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

近年来，人工智能图像生成技术发展迅猛，AI 生成图像已广泛应用于数字文创、游戏美术、影视制作、虚拟仿真、产品设计等领域，极大降低了图像内容创作门槛[1]。但与此同时，生成图像的恶意滥用问题日益突出，不法分子利用高度逼真的 AI 伪造图像传播虚假信息、制造网络谣言、伪造证据、实施网络诈骗，严重扰乱网络秩序、损害公众权益甚至引发社会信任危机[2] [3]。因此，精准区分真实拍摄图像与 AI 生成图像，对网络内容治理、数字媒体取证、信息安全防护、社会舆论稳定均具有十分重要的现实意义[4]。

面向 AI 生成图像真伪识别任务，目前国内外研究者已提出大量检测算法[5] [6]。早期检测方案大多依赖人工设计判别特征，主流思路分为空域纹理统计特征提取与频域生成伪影分析两类。国内研究方面，Zhong 等人[7]提出的 PatchCraft 方法通过“Smash&Reconstruction”预处理增强纹理块，并利用富纹理与贫纹理区域间的像素相关性对比完成真伪判别，在 17 种生成模型基准上优于现有基线；但该方法对纹理线索依赖较强，图像受压缩、噪声干扰时检测鲁棒性仍显不足。在注意力机制方面，Woo 等人[8]提出的 CBAM 以通道与空间注意力串行组合强化关键特征响应，有效提升细微特征感知能力；但其顺序串联结构对多尺度、多层级伪造特征的协同建模能力有限，难以充分挖掘生成图像的频域与纹理复合差异。

近年来，以 GAN、扩散模型为代表的图像生成技术迭代速度极快，Stable Diffusion、Wukong、ADM 等模型产出的合成图像视觉逼真度持续提升，人眼难以直接分辨真假[9]。在此背景下，仅依靠单一空域或频域特征的传统检测方法，面对训练阶段未见过的新型生成模型时判别能力会出现显著下降，难以满足真实场景下通用鉴伪需求。

为突破单一特征检测方法的局限性，基于深度学习的多模态特征融合检测方法逐步成为该领域的主流研究方向。此类方法能够同步挖掘图像空域纹理、频域伪影、梯度分布等多维度伪造特征，通过多特征互补的方式弥补单一特征判别存在的缺陷，大幅提升模型对高逼真 AI 合成图像的鉴别能力与场景适应性[10]。Zhou 等人[11]的研究便通过融合多维度图像特征，有效丰富了模型的判别依据，显著优化了 AI 图像真伪检测精度。但现有多模态融合检测方法仍存在明显短板，多数方案采用固定的特征融合策略，无法根据不同生成模型、不同后处理扰动的图像特征差异，动态适配多模态特征间的关联权重，难以精准捕捉复杂场景下的细微伪造痕迹[12]。为改善模型性能，诸多研究针对性引入优化模块赋能检测模型，其中 Frank 等人[13]、Zhang 等人[14]聚焦频域特征优化，通过强化图像频域伪影挖掘提升模型抗干扰能力，但该类方法过度依赖频域单一维度特征，面对持续迭代优化的 AI 生成模型，伪造图像频域瑕疵不断被优化，检测泛化性能持续受限[15] [16]。Tan 等人[17]提出的梯度特征检测方案，虽依托图像梯度特性有效提升了模型跨域检测能力，但模型性能高度依赖预训练权重质量与足量训练数据支撑，在小样本场景、复杂扰动场景下检测稳定性不足，难以适配当前多样化、高逼真、强伪装的 AI 图像伪造场景，整体通用检测能力仍有待进一步突破。

尽管现有研究已经取得了一定进展，但目前的方法仍然面临两个关键问题。第一个问题是，现有方法在特征提取阶段使用的特征缺少细腻度的刻画，难以充分挖掘和放大图像中与伪造痕迹密切相关的高频噪声、边缘纹理及频域变换等关键特征，这直接限制了模型对复杂生成伪影的感知能力[18] [19]。第二个问题在于，传统注意力机制中的线性投影层表达能力有限，很难有效捕获图像特征之间的高阶非线性关系，这就导致模型对那些复杂伪造痕迹的判别能力不足。

针对第一个挑战，本文一方面提取高频噪声、边缘梯度及频域变换等多类判别特征，另一方面对提取到的多类特征进行自适应增强，以放大与伪造痕迹密切相关的关键信息，弥补单一特征刻画细腻度不足的问题。针对第二个挑战，本文引入深度残差环架构。具体设计如下：在标准多头注意力计算之后，串联多个残差环块；每个残差环块内部依次包含层归一化、非线性前馈映射和 GELU 激活函数，并通过局部残差连接将精炼结果累加回注意力输出。通过累积残差连接，该架构能够实现深度特征的重用，并让梯度在多层之间平滑流动。需要强调的是，该结构并不替代传统多头注意力中的 Q/K/V 线性投影，而是在注意力计算完成后，对注意力输出进行进一步的非线性精炼与高阶关系建模。

本文的贡献总结如下。

面向 AI 生成图像检测任务，本文设计了多模态特征融合框架 R^2-MFAINet。该框架融合通道——空间注意力机制、深度非线性投影环与残差循环架构，从整体上增强了模型对复杂生成图像的识别能力。

设计了结合通道——空间注意力自适应残差模块(CSAARM)，在轻量化残差网络中首先提取高频噪声、边缘纹理及频域特征等多类判别线索，进而利用通道－空间双重注意力机制对关键特征进行自适应加权增强，有效突出与伪造痕迹高度相关的判别信息。

设计了以残差环为核心的深度特征交互机制：在标准多头注意力计算之后引入残差环，对注意力输出进行迭代式非线性后处理，用于捕获特征间的高阶非线性关系。该机制不改变传统注意力中的 Q/K/V 线性投影过程，而是在其输出端进一步增强特征交互与非线性表达能力。

本方法在 GenImage 基准测试上取得了优异的检测性能，精确率、召回率和 F1 分数均表现突出，表明其在检测精度、判别平衡性及泛化能力上优势明显，可有效应对当前主流生成模型产生的多样化伪造图像。

2. 模型结构

本文的任务是针对 AI 生成图像检测中特征提取不充分与高阶非线性关系建模能力不足的问题，提出

一种高效的多特征多注意力集成网络模型。我们采用的模型结构如图 1(1)所示，主要由多特征提取、通道－空间注意力自适应残差模块(CSAARM)、多残差环注意力模块、深度融合模块构成。其整体架构如图 1 所示。

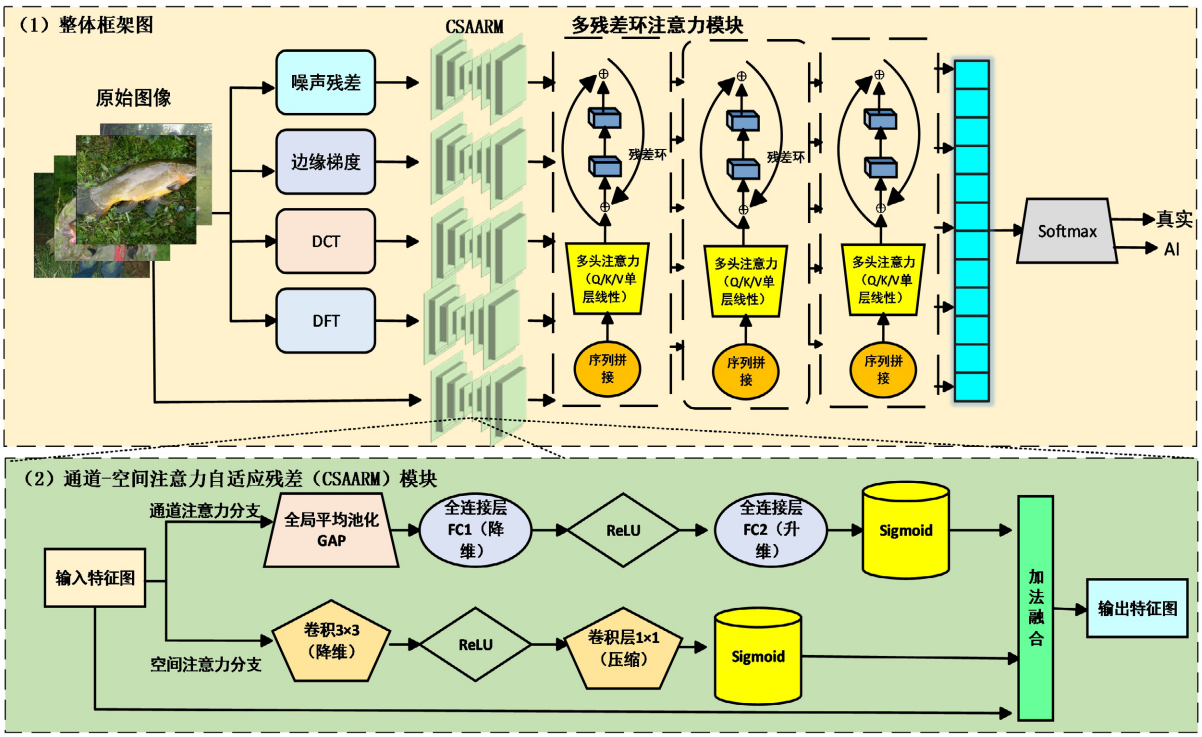


Figure 1. Overall framework of the proposed method
图 1. 整体框架图

在多特征提取阶段，模型从高频噪声、边缘梯度和频域变换三个维度分别提取判别特征：采用均值滤波与原始图像差分获取高频噪声成分，利用 Sobel 算子计算水平与垂直方向的边缘梯度响应，并通过二维离散傅里叶变换和离散余弦变换将图像转换至频率域以提取全局频谱特征，从而从多个维度全面捕捉真实图像与 AI 生成图像之间的本质差异。

在特征增强阶段，通道－空间注意力自适应残差模块(CSAARM)采用并联的双注意力分支结构，对多特征提取阶段得到的特征图进行自适应优化。通道注意力分支通过挤压－激励操作对全局上下文进行建模，自适应重标定各通道的响应权重；空间注意力分支则通过卷积层压缩通道并生成空间注意力图，聚焦伪造痕迹显著的关键区域。两分支输出与原始输入特征通过残差连接相加，在保留底层基础特征的同时强化关键判别特征的表达。

在特征交互阶段，多残差环注意力模块(R^2-MHA)在标准多头注意力计算完成后，将多个非线性前馈块与 GELU 激活函数交替堆叠，并通过局部残差连接构成残差环，对注意力输出进行迭代式深度非线性后处理。该模块不替代传统多头注意力中的 Q/K/V 线性投影层，而是在注意力输出端进一步深化特征间的交互，从而有效捕获多源特征之间的高阶非线性耦合关系。

在融合分类阶段，深度融合模块基于注意力机制对各分支提取的特征向量进行自适应加权融合，动态调整不同模态特征对最终判别结果的贡献比例，强化判别性强的特征来源并抑制冗余干扰信息，融合后的特征输入分类器完成真伪判别。

2.1. 多特征提取

图像特征通常包括空间域特征(如原始像素、边缘纹理和噪声分布)与频域特征(如傅里叶变换和离散余弦变换分量)。针对生成模型所生成图像在这些特征维度上的细微缺陷，本文构建了一个多模态特征提取框架，将原始图像、噪声信息、像素差异信息、傅里叶变换及离散余弦变换特征共同作为网络输入，从多个维度全面捕捉真实图像与 AI 生成图像之间的本质差异。

针对 AI 生成图像检测任务，本文模型从高频噪声、边缘梯度和频域变换三个维度提取判别特征。在噪声特征提取方面，本文采用窗口大小为 5×5 的均值滤波对原始图像进行平滑处理，得到低频分量后将其与原始图像相减，从而分离出高频噪声成分，有效保留生成模型在图像生成过程中留下的噪声模式特征。在边缘梯度特征方面，考虑到生成模型在上采样过程中采用的插值与反卷积操作容易导致边缘区域模糊和像素值异常，本文采用 Sobel 算子分别计算图像在水平与垂直方向上的梯度，并通过加权融合得到综合的边缘梯度特征图，以捕捉边缘区域的细微失真。在频域特征方面，本文分别引入二维离散傅里叶变换和离散余弦变换，将图像从空间域转换至频率域以提取全局频谱特征，其中傅里叶变换用于获取完整的频域复数信息，离散余弦变换则凭借更好的能量集中特性突出图像的整体轮廓与纹理分布，通过频谱中心化操作使模型能够更有效地学习不同频率成分的分布差异。上述多模态特征共同构成模型的输入，从多个角度增强了对真实图像与生成图像的本质区分能力。

2.2. 通道–空间注意力自适应残差特征增强模块

传统的卷积神经网络在处理多尺度特征时，受限于固定大小的感受野，难以同时捕获局部细节与全局上下文信息；在处理多模态特征时，其静态的卷积核权重无法对不同模态特征进行差异化建模，导致特征融合过程中关键信息被背景噪声淹没，尤其在面对复杂生成伪影时特征表达能力明显不足。为增强模型对关键特征的感知能力，本文提出了通道–空间注意力自适应残差模块(CSAARM)，通过双重注意力机制实现自适应的特征优化。该模块采用层级注意力架构，其中通道注意力分支通过挤压–激励操作对输入特征进行全局上下文建模，学习各通道间的依赖关系并自适应重标定通道权重，以强化包含判别信息的通道响应；空间注意力分支则在通道压缩的基础上，通过空间维度的注意力映射聚焦于伪造痕迹显著的区域，增强模型对局部异常模式的感知能力。两者并联融合后与原始输入通过残差连接相加，在保留原始特征的同时有效增强了关键判别特征的表征能力，使模型对生成模型引入的细微伪影具备更强的判别力。其工作流程如图 1(2)所示。

本模块采用双通道并行分支结构处理输入特征图 $F_{in}\in\mathbb{R}^{C\times H\times W}$。输入特征图同时分流至通道注意力分支与空间注意力分支，两条分支独立运算得到各自注意力权重，最终将两组权重与原始输入特征共同送入加法融合单元，输出增强后的特征图 F_{out}。整体过程可表示为：

$$F_{out}=F_{in}+A_c\left(F_{in}\right)\odot F_{in}+A_s\left(F_{in}\right)\odot F_{in}$$

其中 $\odot$ 表示逐元素相乘，A_c 与 A_s 分别为通道与空间分支生成的注意力权重。

在通道注意力分支中，输入特征图首先经过全局平均池化(GAP)完成特征聚合，得到 $z\in\mathbb{R}^C$；随后依次通过降维全连接层 FC1、ReLU 激活、升维全连接层 FC2，还原至原始通道数，最后经 Sigmoid 归一化生成通道注意力权重：

$$A_c=\sigma\left(W_2\cdot\delta\left(W_1\cdot z\right)\right)$$

其中 δ 为 ReLU 函数，σ 为 Sigmoid 函数，W_1,W_2 为全连接层权重。该分支可使网络自主判别不同通道的重要程度，着重保留对真伪识别具备高判别价值的通道特征。

空间注意力分支同步处理输入特征图，先经 3 × 3 卷积层压缩维度并提取局部纹理特征，经 ReLU 激活后送入 1 × 1 卷积层进一步整合通道信息，最后经 Sigmoid 输出单通道空间注意力权重：

$$A_s = \sigma\left(\mathrm{Conv}_{1\times1}\left(\delta\left(\mathrm{Conv}_{3\times3}\left(F_{in}\right)\right)\right)\right)$$

该分支聚焦图像空间区域的细节差异，能够精准定位包含伪造痕迹的关键区域，对细微局部特征的捕捉能力优于全局池化方式。

不同于传统串行叠加的注意力架构，本文并行双分支设计可避免通道与空间注意力运算中的相互干扰。两条分支输出的注意力权重与原始特征经加法融合，依靠可学习参数动态平衡两类注意力的贡献占比：

$$F_{out} = F_{in} + \lambda_c \cdot A_c \odot F_{in} + \lambda_s \cdot A_s \odot F_{in}$$

其中 λ_c，λ_s 为可学习权重参数。同时模块增设残差直连结构，原始输入特征可直接传递至融合端，完整保留底层基础特征；直连通路还能有效缓解深度网络训练时的梯度消失问题，保障模型收敛稳定性。

2.3. 基于残差环的多头注意力特征交互模块

经前述通道－空间注意力自适应残差模块增强后，特征图在关键判别区域与通道维度上已得到初步强化。然而，该阶段仍以单路特征提取为主，不同模态特征之间的高阶非线性耦合关系尚未被充分挖掘，因此需要在增强特征的基础上，进一步提取多源特征之间的深层交互信息，以更全面地捕捉真实图像与 AI 生成图像在频域、纹理、噪声等多维度上的本质差异，从而为最终的真伪鉴别提供更具判别力的融合特征。

图像特征往往具有高度复杂的非线性结构，例如纹理信息在空间分布上呈现出自相似性与多尺度重复模式，边缘区域蕴含着方向性与局部突变特性，频域分量则表现出能量集中性与相位相关性等多重属性。这些丰富的非线性特征难以通过简单的线性变换充分建模，需要更具表达能力的特征交互机制来捕捉其内在关联。在处理多模态特征融合任务时，传统多头注意力(Multi-Head Attention, MHA)中的线性变换层表达能力常常成为性能瓶颈，其单一的线性映射难以揭示多源特征之间的复杂耦合关系。为克服这一限制，并更高效地融合图像、噪声、频域等多源异构特征，本文提出一种残差环增强的多头注意力机制(Residual Ring-Enhanced Multi-Head Attention, R^2-MHA)。该机制通过引入循环残差结构替代传统线性投影，以多层非线性变换迭代深化特征间的交互，从而增强模型对复杂非线性模式的建模能力。

设输入特征序列为 $X \in \mathbb{R}^{B\times L\times D}$，其中 B 为批量大小，L 为序列长度，D 为特征维度。标准多头注意力首先使用三个线性变换生成查询、键、值：

$$Q = XW_Q$$

$$K = XW_K$$

$$V = XW_V$$

其中 W_Q，W_K，$W_V \in \mathbb{R}^{D\times D}$，将 Q，K，V 拆分成为 H 个头，第 i 个头执行缩放点积注意力：

$$Attn_i = \mathrm{softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_h}}\right)V_i$$

拼接各头并经输出投影得到：

$$Z = \mathrm{Concat}\left(Attn_1, \cdots, Attn_H\right)W_O$$

为进一步增强特征交互，在 Z 后串联 R 个残差环块，每个块包含层归一化、膨胀前馈网络和局部残差连接。第 r 块的输入输出为：

$$Z^{(0)} = Z, Z^{(r)} = Z^{(r-1)} + \mathrm{FFN}\left(\mathrm{LayerNorm}\left(Z^{(r-1)}\right)\right)$$

$$\mathrm{FFN}(u) = W_2\,\mathrm{GELU}\left(W_1 u + b_1\right) + b_2$$

整个残差环的最终输出为 $Y = Z^{(R)}$。即所有环块更新量累加至原始注意力输出，形成“环状”信息流。该结构在保证计算效率的同时深化了特征融合，缓解了梯度消失问题。实验表明其性能优于简单平均融合。

这里对“残差环”的命名依据作进一步说明。首先，从残差角度看，每个残差环块均采用局部残差连接，且整个结构的最终输出可视为原始注意力输出 Z 与多级非线性更新量之和，即：

$$Y = Z + \sum r\Delta Z(r)$$

$$\Delta Z(r) = \mathrm{FFN}\left(\mathrm{LayerNorm}\left(Z(r-1)\right)\right)$$

这种累加式残差结构使原始注意力信息能够直接传递到最终输出，避免深度非线性变换造成的信息丢失。其次，从“环”的角度看，该结构并非替代注意力中的 Q/K/V 线性投影，也不是在注意力权重计算内部形成闭环，而是在注意力输出之后，让多个残差环块反复作用于同一特征表示：每一轮精炼的输出都会通过残差连接回写到当前特征中，并作为下一轮精炼的输入。因此，信息流呈现为“注意力输出–非线性变换–残差回写–再变换”的迭代过程。由于每一轮更新都围绕同一注意力输出进行累积式回写，形成类似环状反复精炼的信息通路，故将其称为“残差环”。这一命名强调的是注意力输出之后的深度非线性后处理机制，而非对传统线性投影的替代。

2.4. 多特征融合与加权分类判别

在完成高频噪声、边缘梯度和频域变换等多维度特征的提取后，模型需对各特征来源进行有效融合，以充分利用异构特征间的互补信息。本文采用基于注意力机制的特征融合策略，自适应地调整不同特征模态对最终判别结果的贡献权重。具体地，设各分支提取的特征向量分别为特征模态总数。首先将所有特征向量沿通道维度拼接为多模态特征序列 $f_1, f_2, \cdots, f_k$，随后通过注意力模块计算各模态的权重系数：

$$\alpha_k = \frac{\exp\left(w_k^{\mathrm{T}} f_k\right)}{\sum_{j=1}^{K} \exp\left(w_j^{\mathrm{T}} f_j\right)}$$

$$F_{fuse} = \sum_{k=1}^{K} \alpha_k \cdot f_k$$

该融合策略使模型能够根据输入图像的具体特性动态调整各模态特征的贡献比例，强化判别性强的特征来源，抑制冗余或干扰信息，从而为图像真实性检测任务提供更全面、更精确的特征支撑。融合后的特征向量 F_{fuse} 进一步输入至由全连接层与 Softmax 激活函数构成的分类器中，完成真实图像与 AI 生成图像的二分类判别。分类器通过最小化交叉熵损失函数优化网络参数，使模型在训练过程中逐步学习到更具判别性的融合特征表达，从而实现对各类生成模型所产生伪造图像的有效鉴别。

2.5. 损失函数

为实现多分支网络协同训练、充分挖掘空域、频域、梯度等多模态特征间互补信息，本文构建多层次复合监督损失函数。该损失函数包含双监督分支，分别为各单模态独立输出预测产生的交叉熵损失累

加项、多特征融合模块最终输出对应的交叉熵损失项，整体损失由两项等权重加权求和得到，计算公式定义如下：

$$\mathcal{L}_{total} = \mathcal{L}_{branch} + \mathcal{L}_{fusion} = \sum_{i=1}^{M} \mathcal{L}_{CE}\left(P_i, Y\right) + \mathcal{L}_{CE}\left(P_{fusion}, Y\right)$$

式中，L_{total} 代表整体复合损失；M 为多模态分支总数量；P_i 表示第 i 个单模态分支输出的真伪预测概率，P_{fusion} 为特征融合模块输出的综合预测结果；Y 为样本真实标签；L_{CE} 代表标准二分类交叉熵损失函数。模型训练阶段通过反向传播同步回传总损失梯度，统一更新全部单模态特征提取分支与融合模块参数。该分层监督策略可向每一路模态分支独立传递梯度约束，有效训练各分支特征提取能力、抑制单一模态特征退化；同时融合模块在全局损失引导下挖掘不同模态间差异化伪造特征，充分利用多维度信息互补优势，最终提升模型对各类 AI 生成图像的泛化检测精度。

3. 仿真实验与结果分析

3.1. 数据准备

本文采用 GenImage 数据集[20]中的部分数据集作为实验评测基准。该数据集共包含 84,000 张图像，其中真实图像与 AI 生成图像各 42,000 张。为了全面评估模型的检测性能，本文将整个数据集按 8:2 的比例随机划分为训练集和测试集，其中训练集占图像总数的 80%，测试集占 20%。在训练集内部，进一步按照 8:2 的比例随机划分出训练子集(占训练集的 80%，即整体的 64%)和验证子集(占训练集的 20%，即整体的 16%)，用于模型训练过程中的参数更新与超参数调优。测试集在模型训练和验证阶段完全不可见，仅用于最终性能评估。所有图像均统一缩放到 256 × 256 分辨率，并采用相同的数据预处理流程(包括均值滤波、Sobel 梯度计算、DFT/DCT 变换等)，以确保实验的公平性与可复现性。

3.2. 评价指标

本文采用精确率(Precision)、召回率(Recall)和 F1 分数作为主要评价指标，以全面评估模型在图像篡改检测任务中的性能。对于二分类任务，各指标定义如下：

精确率(Precision)：TP/(TP + FP)

召回率(Recall)：TP/(TP + FN)

F_1 分数：2 × (Precision × Recall)/(Precision + Recall)

其中 TP、TN、FP、FN 分别表示真阳性、真阴性、假阳性、假阴性。所有实验均在相同硬件环境下进行，保证结果可比性。

3.3. 实验结果

为全面评估 R^2-MFAINet 在 AI 生成图像检测任务中的性能，本节从检测精度、泛化能力、模块有效性和计算复杂度四个维度展开实验。在具体实验开展之前，首先对模型的计算复杂度进行简要分析。R^2-MFAINet 的计算开销主要来源于多模态特征提取分支与残差环注意力模块：前者并行处理原始图像、噪声、DCT、DFT 和边缘梯度五类输入，其计算量随模态数量增加而近似线性增长；后者通过残差环迭代实现跨模态特征精炼，其复杂度与残差环堆叠次数及多头注意力头数呈正相关。根据后续消融实验，当残差环堆叠 3 次、注意力头数为 8 时，模型在验证集上达到 99.67%的最优准确率，同时未因迭代过深或注意力头数过多而引入明显计算冗余，说明该配置在检测精度与计算开销之间取得了较好平衡。基于这一最优配置，下文依次开展对比实验、量化分析、泛化实验、消融实验与可解释性分析，以系统验证模型的综合性能。

3.3.1. 对比实验

为充分验证本文提出的 R²-MFAINet 模型在区分真实图像与 AI 生成图像任务上的优越性，本文采用 GenImage 数据集中四类不同的生成图像子集(ADM, BigGAN, GLIDE, Wukong)进行实验。实验设置统一的数据预处理流程、输入分辨率与评价指标，构建多层次对比基线：第一类为传统经典检测算法，包含基于 DFT 频谱图的检测方法[14]、仅依靠真实图像学习噪声特征的检测方法[2]、基于 DCT 频域分析的伪造图像识别方法[13]、融合局部与全局特征的通用 AI 图像检测方法[10]；第二类为 2024 年基于 GenImage 数据集的主流通用检测模型 IPD-Net [21]；第三类为 2025 年 ICCV 顶会提出的当前 SOTA 检测方法 LOTA [22]。各模型在相同实验环境下的精确率、召回率与 F1 分数对比结果如表 1 所示。

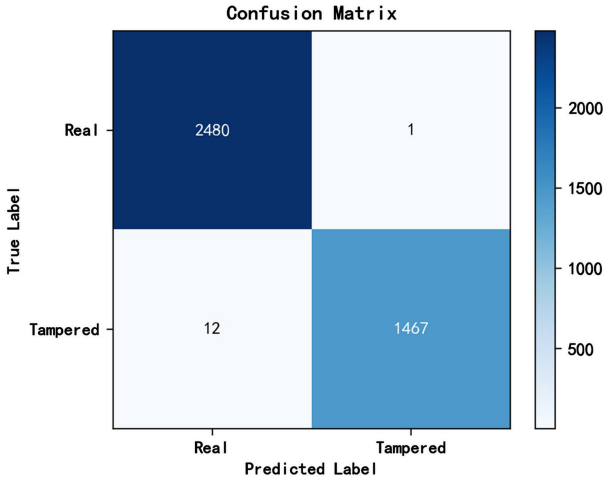
Table 1. Comparison of experimental results
表 1. 对比实验结果

模型	Precision score	Recall score	F_1 分数
文献[14]	0.9199	0.9159	0.9179
文献[2]	0.9947	0.9718	0.9841
文献[13]	0.9889	0.8557	0.9175
文献[10]	0.9476	0.9536	0.9507
文献[21]	0.8743	0.8743	0.9628
文献[22]	0.9826	0.9779	0.9802
R²-MFAINet	**0.9967**	**0.9919**	**0.9956**

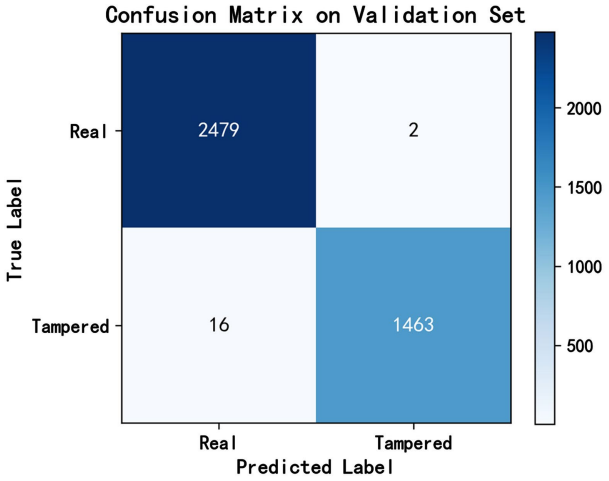
本文模型在精确率、召回率和 F1 上均取得最佳结果，表明其通过多模态融合与残差环注意力机制能更精准捕获生成图像的判别性伪影。

3.3.2. 量化分析

本实验采用的是二分类模型。验证集和测试集各包含 3960 张图像，每张图像的真实标签都是已知的。模型做出预测之后，按照真实标签和预测标签的 4 种组合分别统计样本数量，得到两个混淆矩阵，如图 2 所示。



(a) 测试集混淆矩阵



(b) 验证集混淆矩阵

Figure 2. Confusion matrix
图 2. 混淆矩阵

测试集上混淆矩阵显示：准确率 99.67%，漏报(FN)为主(测试集 12 例)，误报(FP)极少(1 例)，符合低误报率的实际需求，且测试集指标略优于验证集，无过拟合。

3.3.3. 泛化实验

为检验泛化性，我们采用 GenImage 中未经训练触及的 Midjourney、Stable Diffusion 和 VQDM 三个子集作为测试集，将已训练好的 R^2-MFAINet 直接用于推理。准确率结果如表 2 所示，该测试集与训练集完全无重叠，保证了评估的可靠性。

Table 2. Generalization experiment results
表 2. 泛化实验结果

	Precision score	Recall score	F_1 分数
Midjourney	0.8310	0.7125	0.6950
stablediffusion	0.8881	0.9063	0.7350
VQDM	0.8010	0.8003	0.8253

模型在三个未见生成器生成的图像集均取得了较为理想的检测效果。其中，在 VQDM 数据集上表现最佳，F1 分数达到 0.8253，精确率与召回率高度均衡(0.8010 与 0.8003)，表明模型对该类生成图像的判别最为稳定可靠；在 Stable Diffusion 上精确率达 0.8881、召回率达 0.9063，展现出优异的真伪分辨能力；在 Midjourney 上召回率相对偏低(0.7125)，可能源于其生成图像质量更高、伪造痕迹更为隐蔽，但整体仍达到可接受的检测水平。三个未见生成器上的 F1 分数均达到 0.6950 以上，充分验证了模型在面对训练阶段完全未接触的生成模型时，仍具备良好的跨生成器泛化能力。

3.3.4. 消融实验

为全面验证 R^2-MFAINet 各组成部分的有效性，我们设计了两个层次的消融实验：模态分支消融和残差环结构消融。前者考察不同输入模态的贡献，后者则深入探究残差环内部机制对最终性能的影响。

为深入探究残差环的迭代次数及内部注意力配置对性能的影响，我们在全部五种模态输入的基准下，固定其他超参数，分别调整环内迭代次数以及多头注意力的头数，结果如表 3 所示。

Table 3. Ablation experimental results
表 3. 消融实验结果

残差环堆叠次数	注意力头数为 4	注意力头数为 8	注意力头数为 16
0	94.11%	95.96%	95.01%
1	96.64%	97.49%	97.33%
2	99.60%	99.48%	97.12%
3	97.98%	**99.67%**	96.55%
4	96.11%	98.99%	94.01%

当固定堆叠次数不变时，多头注意力头数对性能的影响表现出明显的阈值效应：头数为 8 时在各堆叠次数下均取得最优结果(峰值 99.67%)，而头数为 16 时性能反而出现下降，表明过多的注意力头数可能引入冗余或噪声干扰，削弱了模型的判别聚焦能力。

在固定注意力头数不变的情况下，残差环堆叠次数对性能的影响呈现明显的先升后降趋势。以注意力头数为 8 的条件为例，堆叠次数从 0 逐步增加到 2 时，准确率从 95.96%持续提升至 99.67%；而继续增加至 3 次和 4 次时，性能分别回落至 99.48%和 98.99%，表明适量的迭代精炼有助于深化特征交互，但过深的堆叠反而可能导致过度平滑或过拟合。

综合来看，最优配置为堆叠 2 次与 8 头注意力(99.67%)，此时模型在特征交互深度和并行子空间分割之间取得了最佳平衡。值得注意的是，即便在无残差环堆叠(0 次)时，8 头配置仍能达到 95.96%的准确率，验证了多头注意力自身对多模态融合的有效性；而引入残差环后性能进一步提升 3.71 个百分点，充分证明了残差环通过非线性迭代精炼确实能够有效捕获跨模态特征间的高阶耦合关系。

我们固定残差环结构为 3，依次移除或添加各模态分支，在验证集上评估精确率，结果如表 4 所示。

Table 4. Ablation results of different modality branches
表 4. 各模态分支消融实验结果

原始图像	噪声分支	DCT 分支	DFT 分支	边缘梯度	精确率
√	×	×	×	×	95.96%
√	×	√	√	√	99.39%
√	√	√	√	×	99.47%
√	√	√	√	√	99.67%

从表中可以看出，仅使用原始图像时准确率为 95.96%，说明单模态信息有限。在加入 DCT、DFT 和边缘梯度分支后，准确率跃升至 99.39%，表明频域与边缘线索提供了强判别性特征。进一步引入噪声分支并移除边缘梯度，准确率提升至 99.47%，证实噪声域同样包含有效伪造痕迹。最终，同时融合全部五个模态，模型达到最优的 99.67%，相比缺噪声和缺边缘的设置分别提升 0.28%和 0.20%，充分证明各模态特征互为补充，联合使用能最大化检测性能。

3.3.5. 可解释性分析

为了探究 R^2-MFAINet 各个输入模态对最终决策的贡献有多大，采用了特征屏蔽法：依次把某一模态的输入置为零张量，重新计算模型在验证集上的 F1 分数，观察它下降了多少；下降越多，说明这个模态越重要，结果如图 3 所示。

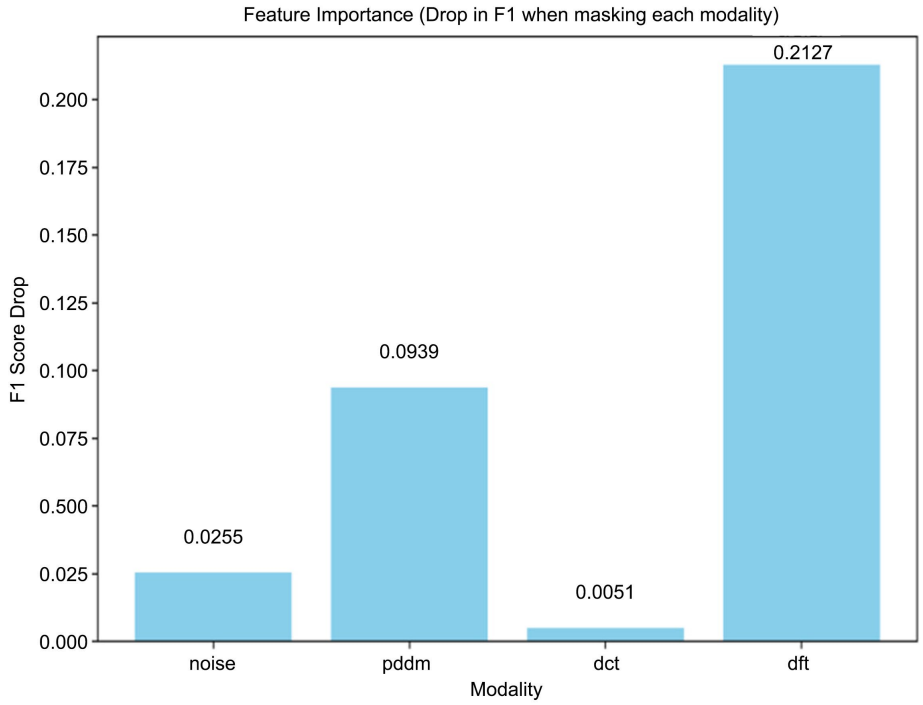


Figure 3. Interpretability analysis
图 3. 可解释分析

DFT 被屏蔽后 F1 下降最多(0.2127)，表明频域相位信息最为关键；边缘梯度(PDDM)次之(下降 0.0939)；Noise 下降 0.0255；DCT 下降最少(0.0051)。整体上，多模态融合使模型具备鲁棒性，即使屏蔽任一分支仍保持较高性能。

4. 结论

针对现有 AI 生成图像检测方法中单模态特征判别能力不足、传统注意力对高阶关系建模有限、深度网络特征传递效率低等问题，本文提出了残差环注意力驱动的多模态融合网络 R^2-MFAINet。该方法从特征提取、注意力优化和深度架构三个层面进行了系统性创新。

具体地，构建了涵盖原始图像、高频噪声、边缘梯度、DFT 和 DCT 的五模态特征框架，从空域、频域和梯度域多维度捕捉真实与生成图像的本质差异。设计了通道－空间注意力自适应残差模块(CSAARM)，通过并联双重注意力与可学习融合策略增强关键特征感知，并以轻量化残差结构缓解梯度消失。进一步提出了残差环增强的多头注意力机制，在标准多头注意力计算之后引入残差环，通过多层非线性前馈变换与残差连接对注意力输出进行迭代式深度后处理，显著提升跨模态交互的表示复杂度。该机制不替代传统注意力中的 Q/K/V 线性投影，而是在注意力输出端进一步增强高阶非线性建模能力。

在 GenImage 基准上的实验表明：完整模型在四种主流生成模型的混合测试集上达到 99.67%的精确率；在 Midjourney、Stable Diffusion 和 VQDM 三个未见生成器上的泛化实验中，模型分别取得 0.6950、0.7350 和 0.8253 的 F1 分数，验证了跨生成器的强泛化能力；与仅使用单模态输入的方法相比，完整模型的检测精度高出 3.71 个百分点。消融实验证实各模态特征均有正向贡献，且模型对特征冗余具有良好鲁棒性。收敛性分析显示，多模态输入使模型在 3~4 个 epoch 内准确率超过 98%，训练过程稳定且无明显过拟合现象。上述结果充分证明了本方法在检测精度、判别平衡性、收敛效率及跨模型泛化能力上的显著优势。

参考文献

[1] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., *et al*. (2015) Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. 2015 *International Conference on Machine Learning*, Lille, 6-11 July 2015, 2256-2265.

[2] Bhowal, A., Naskar, R. and Neogy, S. (2025) Multi-Approach Survey and In-Depth Analysis of Image Forgery Detection Techniques. *The Visual Computer*, **41**, 9977-10035. https://doi.org/10.1007/s00371-025-04019-z

[3] Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I. and Gavves, E. (2025) Any-Resolution AI-Generated Image Detection by Spectral Learning. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 18706-18717. https://doi.org/10.1109/cvpr52734.2025.01743

[4] Cozzolino, D., Poggi, G., Corvi, R., Nießner, M. and Verdoliva, L. (2024) Raising the Bar of AI-Generated Image Detection with Clip. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 17-18 June 2024, 4356-4366. https://doi.org/10.1109/cvprw63382.2024.00439

[5] Zhang, H., He, Q., Bi, X., Li, W., Liu, B. and Xiao, B. (2025) Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 23828-23837. https://doi.org/10.1109/cvpr52734.2025.02219

[6] Cheng, S.Y., Lyu, L.J., Wang, Z.T., *et al*. (2025) CO-SPY: Combining Semantic and Pixel Features to Detect Synthetic Images by AI. https://arxiv.org/abs/2503.18286

[7] Zhong, N., Xu, Y.R., Li, S., *et al*. (2023) PatchCraft: Exploring Texture Patch for Efficient AI-Generated Image Detection. https://arxiv.org/abs/2311.12397

[8] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) CBAM: Convolutional Block Attention Module. In: *Lecture Notes in Computer Science*, Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[9] Zhong, N., Chen, H.Y., Xu, Y.R., *et al*. (2025) Beyond Generation: A Diffusion-Based Low-Level Feature Extractor for Detecting AI-Generated Images. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 8258-8268. https://doi.org/10.1109/cvpr52734.2025.00773

[10] He, K.M., Zhang, X.Y., Ren, S.Q. 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

[11] Zhou, P., Han, X., Morariu, V.I. and Davis, L.S. (2018) Learning Rich Features for Image Manipulation Detection. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1053-1061. https://doi.org/10.1109/cvpr.2018.00116

[12] Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K. and Lyu, S. (2022) Fusing Global and Local Features for Generalized AI-Synthesized Image Detection. 2022 *IEEE International Conference on Image Processing* (*ICIP*), Bordeaux, 16-19 October 2022, 3465-3469. https://doi.org/10.1109/icip46576.2022.9897820

[13] Frank, J., Eisenhofer, T., Schönherr, L., *et al*. (2020) Leveraging Frequency Analysis for Deep Fake Image Recognition. 2020 *International Conference on Machine Learning*, Online, 14-16 July 2020, 3247-3258.

[14] Zhang, X., Karaman, S., Chang, S.F., *et al*. (2019) Detecting and Simulating Artifacts in GAN Fake Images. 2019 *IEEE International Workshop on Information Forensics and Security* (*WIFS*), Delft, 9-12 December 2019, 1-6. https://doi.org/10.1109/wifs47025.2019.9035107

[15] Yu, N., Davis, L. and Fritz, M. (2019) Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 7555-7565. https://doi.org/10.1109/iccv.2019.00765

[16] Alabrah, A. (2025) SFTA-Net: A Self-Supervised Approach to Detect Copy-Move and Splicing Forgery to Leverage Triplet Loss, Auxiliary Loss, and Spatial Attention. *PeerJ Computer Science*, **11**, e2803. https://doi.org/10.7717/peerj-cs.2803

[17] Tan, C.C., Zhao, Y., Wei, S.K., Gu, G.H. and Wei, Y.C. (2023) Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 12105-12114. https://doi.org/10.1109/cvpr52729.2023.01165

[18] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/cvpr.2017.243

[19] Chai, L., Bau, D., Lim, S. and Isola, P. (2020) What Makes Fake Images Detectable? Understanding Properties That Generalize. In: *Lecture Notes in Computer Science*, Springer, 103-120. https://doi.org/10.1007/978-3-030-58574-7_7

[20] Zhu, M.J., Chen, H.T., Yan, Q.Y., *et al*. (2023) GenImage: A Million-Scale Benchmark for Detecting AI-Generated

Image. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 77771-77782.

[21] Chen, J.H., Lo, M.T., Liao, H.L., *et al*. (2024) IPD-Net: Detecting AI-Generated Images via Inter-Patch Dependencies. *International Journal of Advanced Computer Science and Applications*, **15**, 1371-1378. https://doi.org/10.14569/ijacsa.2024.01507133

[22] Wang, H.S., Cheng, R.X., Zhang, Y., *et al*. (2025) LOTA: Bit-Planes Guided AI-Generated Image Detection. 2025 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Honolulu, 19-23 October 2025, 13421-13430. https://doi.org/10.1109/iccv51701.2025.01602