融合光影敏感特征及K-A表示定理的AI生成图像鉴别方法

邓 巧[1],姜 林[1,2*],刘乐新[1],唐吕鑫[1],杨英丽[1]
(1. 湖南工商大学人工智能与先进计算学院,湖南长沙 410000; 2. 湘江实验室,湖南长沙 410000)

摘 要: 人工智能(Artificial Intelligence,AI)生成图像技术发展迅猛,高逼真内容对网络安全与社会信任构成重大威胁,而人类自主鉴别准确率仅约59%,接近随机猜测水平. 现有检测方法普遍存在性能有限、跨模型泛化能力不足等问题,尤其无法有效捕捉生成图像中物理光照的不一致性. 为此,本文提出融合光影敏感特征及Kolmogorov-Arnold(K-A)表示定理的特征融合鉴别方法(Light-enhanced Kolmogorov-Arnold Networks,L-KAN). 在红绿蓝三原色(Red、Green、Blue,RGB)语义特征、频域特征和边缘特征的基础上,构建光影敏感特征. 该特征通过整体光照分布、阴影面积及方向和多尺度光照梯度特性,捕捉生成图像中的光照异常. 引入K-A表示定理进行特征融合,通过内外层函数协同作用,在保证特征互补性的同时有效抑制特征冗余. 在3组公开数据集上,与9种先进方法进行对比,所提方法平均分类准确率均有显著提升.

关键词: AI生成图像检测;光影敏感特征;特征融合;Kolmogorov-Arnold表示定理

基金项目: 湖南省教育厅科学研究项目(重点)(No. 22A0441);湘江实验室重大项目(No. 23XJ01003, No.23XJ01009)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)11-4077-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250250

AI-Generated Image Detection Method Integrating Light-Shadow Sensitive Features and Kolmogorov-Arnold Representation Theorem

DENG Qiao[2], JIANG Lin[1,2*], LIU Le-xin[2], TANG Lü-xin[2], YANG Ying-li[2]
(1. *School of AI and Advanced Computing*, *Hunan University of Technology and Business*, *Changsha*, *Hunan* 410000, *China*;
2. *Xiangjiang Laboratory*, *Changsha*, *Hunan* 410000, *China*)

Abstract: The rapid advancement of artificial intelligence (AI) -generated image technologies poses significant threats to cybersecurity and public trust, as human visual detection accuracy remains as low as 59%, close to random guessing. Existing detection methods suffer from limited performance and poor generalization across generative models, particularly struggling to capture physical inconsistencies in illumination. To address this gap, we propose L-KAN (Light-enhanced Kolmogorov-Arnold Networks), a novel detection framework that integrates illumination-sensitive features with the Kolmogorov-Arnold (K-A) representation theorem. Building upon red-green-blue (RGB) semantics, frequency-domain cues, and edge information, we construct physically grounded features that encode global illumination distribution, shadow geometry, and multi-scale illumination gradients to expose lighting inconsistencies in synthetic images. Leveraging the K-A theorem for feature fusion, ours method synergizes inner and outer functions to enhance feature complementarity while suppressing redundancy. Experimental results on three public datasets demonstrate that L-KAN achieves a competitive performance compared with the state of the art methods.

Key words: AI-generated image detection; light-shadow sensitive features; feature fusion; Kolmogorov-Arnold representation theorem

Foundation Item(s): Scientific Research Project of Hunan Provincial Department of Education (No.22A0441); Major Project of Xiangjiang Laboratory (No.23XJ01003, No.23XJ01009)

收稿日期:2025-04-02;录用日期:2025-05-22;责任编辑:覃怀银
*通讯作者:姜林

1 引言

人工智能技术在视觉内容生成领域的飞速发展正在深刻改变数字媒体的创作与传播方式. 特别是生成对抗网络(Generative Adversarial Networks,GAN)[1]和扩散模型(Diffusion Models,DM)[2]等先进算法的出现,使得人工智能(Artificial Intelligence,AI)生成图像的质量与真实性达到了前所未有的高度[3]. 这些技术突破不仅推动创意产业的创新,同时也引发了广泛的社会关注与安全隐患[4]. 在当前信息高度数字化时代,真实内容与AI生成内容的界限日趋模糊,对公众认知、舆论环境和社会信任体系带来了严峻挑战[5~7]. 而根据Nightingale和Farid的实验,即使提供培训和反馈,参与者在识别AI合成人脸时,平均准确率仅提高至59%,只接近随机猜测水平[8].

由于人眼难以识别AI生成图像中的细微差异,迫切需要开发有效的检测方法. 现有的AI生成内容检测方法主要分为空域方法和频域方法两大类[9]. Zhao等人[10]提出成对自一致性学习(Pair-wise self-Consistency Learning,PCL)的空域方法,假设伪造图像中存在源特征的不一致性,通过计算局部块之间的余弦相似度生成一致性灰度图. Qian等人[11]的频域研究表明,频域感知分解模块能在高压缩图像中有效捕捉伪影细节,保持较好的检测性能. 这些研究展示了空域和频域特征在AI生成内容检测中的不同优势与应用. 然而现有方法仍存在显著局限:空域方法主要关注局部纹理和结构特征,难以有效捕捉全局特征分布的不自然性;频域方法虽然能够检测到压缩伪影,但对物理规律性缺乏建模. 这些局限性使得现有方法在检测复杂场景下的AI生成图像时仍有较大提升空间. 本文通过对当前流行的AI生成图像的观察,发现普遍存在亮度分布异常、光照和阴影不一致等问题,如图1所示,而这些问题在现有研究中尚未得到充分探讨.

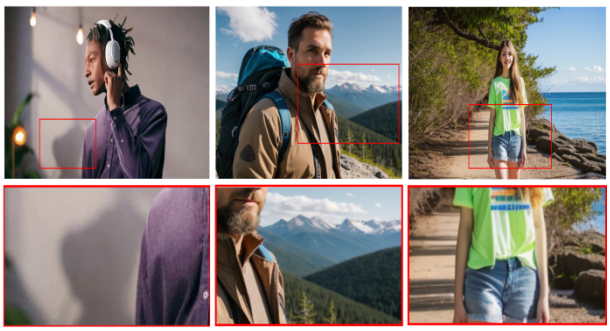


图1 AI生成图像光影异常示例图

面对不断进化的生成模型,多项研究表明,单一特征的检测方法面临着更大的挑战,尤其是在处理多样化的AI生成内容时[12,13]. 单一特征往往不足以捕捉生成图像与真实图像间的细微差异特征,难以全面表达数据中的复杂信息. 早期的线性组合[14,15]、U-Net(Convolutional Networks for Biomedical Image Segmentation)跳跃连接[16]和近年的SENet(Squeeze-and-Excitation Networks)注意力机制[17]等方法取得了一定成功,但仍面临两大关键挑战:融合过程中的噪声干扰问题以及特征冗余与互补性的平衡困境.

针对上述挑战,本文提出融合光影敏感特征及Kolmogorov-Arnold(K-A)表示定理的特征融合鉴别方法. 本文的主要贡献如下:

(1)设计光影敏感特征提取算法,通过分析基础光照统计量(YUV空间亮度分布特征)、阴影物理特征(区域分割与方向分析)和多尺度光照梯度特征(不同尺度下的梯度一致性),有效识别AI生成图像中的光照物理规律缺陷,显著提升跨生成模型场景下的检测性能.

(2)提出基于K-A表示定理的特征融合方法(Kolmogorov-Arnold feature Fusion,KAF),通过内外层函数分别实现特征降维去噪与非线性融合. 采用双层KAF级联架构,使主干特征递进式更新而辅助特征保持不变,在保证特征互补性的同时有效抑制冗余,显著提升模型泛化能力.

2 相关工作

AI生成图像的检测技术是近年来计算机视觉领域的研究热点. 本节将介绍当前AI生成图像检测的主要特征类型,常见的特征融合方法及基于K-A表示定理的相关研究.

2.1 空域特征和频域特征

目前,AI生成图像检测研究主要聚焦于空间域特征和频域特征两大类型[9]. 空间域特征通过分析图像的局部纹理、一致性和几何约束,揭示生成图像与自然图像之间的差异. 频域特征则通过分析图像的频谱分布,捕捉生成过程中引入的伪影和异常.

空域方法聚焦生成图像在像素空间的统计异常与模式表征. Wang等人[18]在2020年的研究中发现,在ProGAN上训练的检测器可以泛化到其他GAN模型,实现了对CNN(Convolutional Neural Network)生成图像的有效检测. 通过精心设计的数据增强和预处理,他们的方法在StyleGAN生成的图像上表现出很高的检测精度. Tan等人[19]在2023年提出了LGrad框架,该方法利用预训练CNN的梯度信息构建通用伪影表示,相比现有方法显著提升了检测性能. 这是首次将梯度用作伪影表示的方法,展现了良好的效果和鲁棒性. Ojha等人[20]的研究表明,传统分类器难以泛化到新生成模型,他们提出使用未经特定训练的特征空间进行分类,如最近邻和线性探测方法,提高了检测的通用性. Tan等

人[21]在2024年的研究中深入分析了上采样操作产生的相邻像素关系(Neighboring Pixel Relationship,NPR),该研究重新评估了CNN生成器架构,聚焦于上采样操作建立泛化的伪造伪影表示. 研究发现,上采样操作不仅产生频率域伪影,还会通过像素间的局部相互依赖关系产生可泛化的伪影模式.

频域方法通过多尺度频率成分分析捕捉生成伪影. Frank等人[22]在2020年的研究中发现,GAN生成的图像在频率域存在明显伪影,这些伪影主要源于上采样操作. 他们利用离散余弦变换(Discrete Cosine Transform,DCT)系数分析,提出了一种有效的检测方法,在未见过的生成模型上仍保持了较高的检测准确率. Qian等人[11]提出了F3-Net,利用频率域分解和局部频率统计挖掘伪造线索. 该方法特别适用于压缩图像和视频,能更好地描述频率域中的细微伪影和压缩错误. 最近,Tan等人[23]通过频率空间学习增强检测泛化性,该方法强制检测器关注高频信息,从而学习源模型无关的特征,使用较少的参数实现优异性能. Durall等人[24]的补充实验进一步揭示,生成器中常用的上采样操作无法复现真实图像的频谱连续性,其发现的频谱特征为频域检测提供了新的物理可解释性指标.

空间域和频域特征虽在AI生成内容检测中取得了显著进展,但在处理图像物理规律约束方面仍存在局限,尤其忽略了高光分布、阴影方向和色温等异常. 因此,本研究在综合分析现有方法的基础上,提出了一种光影敏感特征,进一步揭示生成图像中的物理规律缺陷.

2.2　特征融合

单一特征往往不足以捕捉数据中的复杂信息,特征融合已成为提高模型准确性和效率的必要手段,它作为增强模型表征能力的关键技术,历经从线性组合到非线性架构、从静态设计到动态自适应的发展.

早期研究主要通过数学运算实现浅层特征交互,例如Szegedy等人[14]提出Inception网络通过并行多尺度卷积实现通道堆叠,利用不同感受野的特征增强多尺度感知;He等人[15]设计的ResNet采用残差连接的逐元素相加传递跨层特征有效缓解梯度消失问题. 此类方法虽计算高效,但受限于线性运算对复杂非线性关系的建模能力. Ronneberger等人[16]提出U-Net通过跳跃连接的拼接操作实现编码器-解码器特征传递,保留高分辨率细节;Lin[25]等人开发特征金字塔网络(Feature Pyramid Network,FPN),采用横向连接融合不同层级的语义与空间信息. 近年来,自适应融合方法通过动态机制提升融合效果. Hu等人[16]提出SENet利用通道注意力学习特征权重,突破静态融合局限性;Liu等人[26]通过内省对抗学习实现生成器与判别器的动态反馈,为跨模态特征融合提供了新思路;Zhang等人[27]提出HFFN(Hierarchical Feature Feedback Network)网络通过金字塔结构提升深度图像超分辨率质量,但依赖复杂多尺度计算,难以迁移至生成图像检测任务;Lin等人[28]提出轻量级张量网络,通过张量分解优化异构特征融合效率,为跨领域轻量化设计提供参考;Wilson等人[29]通过深度核架构增强核函数的表达能力,实现了可扩展的特征表示;Sysko-Romańczuk等人[30]基于再生核希尔伯特空间的函数分解理论,构建了可分解的优化框架,为特征融合的轻量化部署和复杂任务处理提供了新的理论基础.

现有的特征融合方法在特定任务中取得了一定的成功,但仍然面临噪声干扰、特征冗余与互补性平衡等关键挑战[31]. 受K-A表示定理启发,本文提出了一种双层函数架构的特征融合框架(KAF).

2.3　K-A表示定理

K-A定理[32]由Kolmogorov和Arnold提出. 该定理表明,任何多元连续函数都可以表示为有限个单变量函数的复合与叠加. 具体来说,对于任意多元连续函数$f\left(x_1,x_2,\cdots,x_n\right)$,可以找到一组单变量函数,使得该多元函数能够通过这些单变量函数的组合来表示. 该定理为复杂的多元函数的表示提供了一种简化的方法,并揭示了多元函数本质的结构:

$$f=\sum_{q=1}^{2n+1}\phi_q\left(\sum_{p=1}^{n}\psi_{q,p}\left(x_p\right)\right) \tag{1}$$

近年来,K-A定理在机器学习中的应用得到了广泛关注和进一步扩展. Schmidt-Hieber[33]探讨了K-A定理在神经网络多层结构中的应用,提出通过修改外函数的平滑性来改善模型的可近似性. Polar等人[34]提出了一种快速稳定的算法,用于构建K-A表示,特别适用于量化输入和连续输入的组合,并在多个公开数据集上优于其他基准方法. Liu等人[35]基于这一理论提出了Kolmogorov-Arnold Networks(KANs),这是一种替代传统MLP(MultiLayer Perceptron)的神经网络架构,通过单变量函数的组合来逼近复杂的多变量函数,显著提高了模型的表示能力. KAN在多个标准任务上实现了与传统深度网络相当甚至更优的性能,同时参数量大幅减少.

受K-A表示定理启发,本文提出了双层函数架构的特征融合框架(KAF),将多特征融合任务分解为两个层次:(1)内层函数(Ψ)实现各特征流的独立处理与归一化,有效降低噪声干扰;(2)外层函数(Φ)通过单变量函数的复合与叠加方式对归一化特征进行非线性融合,构建高判别力表示. 此设计不仅简化了复杂特征空间的映射过程,还通过级联结构实现了主干特征的递进更新与辅助特征的稳定补充,有效平衡了特征互补

性与非冗余性.

3 方法设计

本节按照所提出系统的一般框架展开论述. 该方法主要由4个模块组成:(1)特征提取;(2)特征融合;(3)分类器设计;(4)损失函数优化. 本文所提出框架的流程图如图2所示,重点聚焦于光照特征提取与基于K-A表示定理的特征融合机制.

3.1 特征提取

本文设计四路并行的特征提取模块,从语义、频域、边缘结构和光照4个维度深入分析. RGB语义特征流利用预训练ResNet50捕捉图像整体语义与深层纹理;频域特征流通过傅里叶变换捕捉生成伪影;边缘特征流基于Canny算子提取结构边界信息. 本文重点关注的光影敏感特征通过量化亮度分布、阴影特征及多尺度光照梯度等特征,有效识别生成图像中不符合自然光照规律的异常模式.

通过对真实与AI生成图像的对比分析,本文观察到生成图像在光照分布、阴影特性及光照梯度等方面与真实图像存在明显差异. 如图3的L通道亮度分布、阴影区域分割及梯度场可视化($\alpha=4$)所示,两类图像的光照特性存在差异,为后续特征设计提供了理论依据.

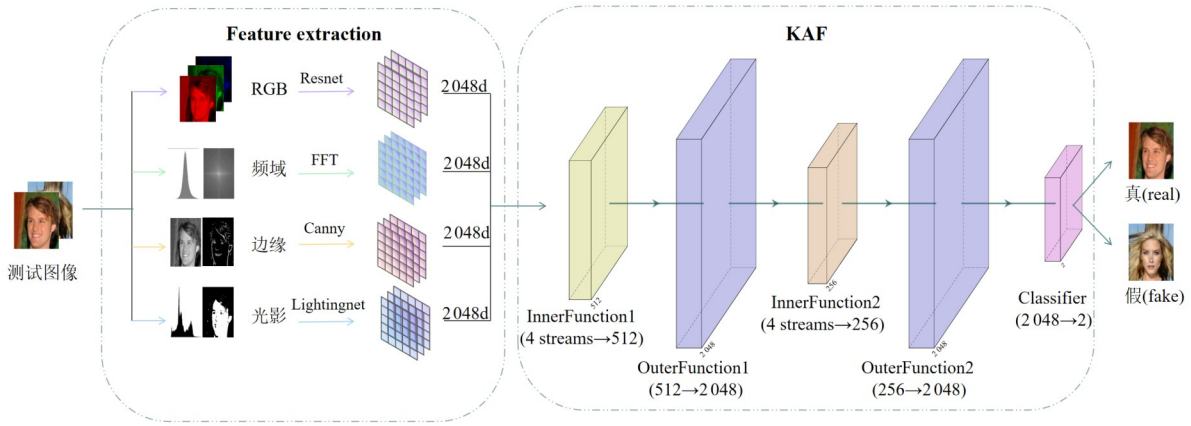
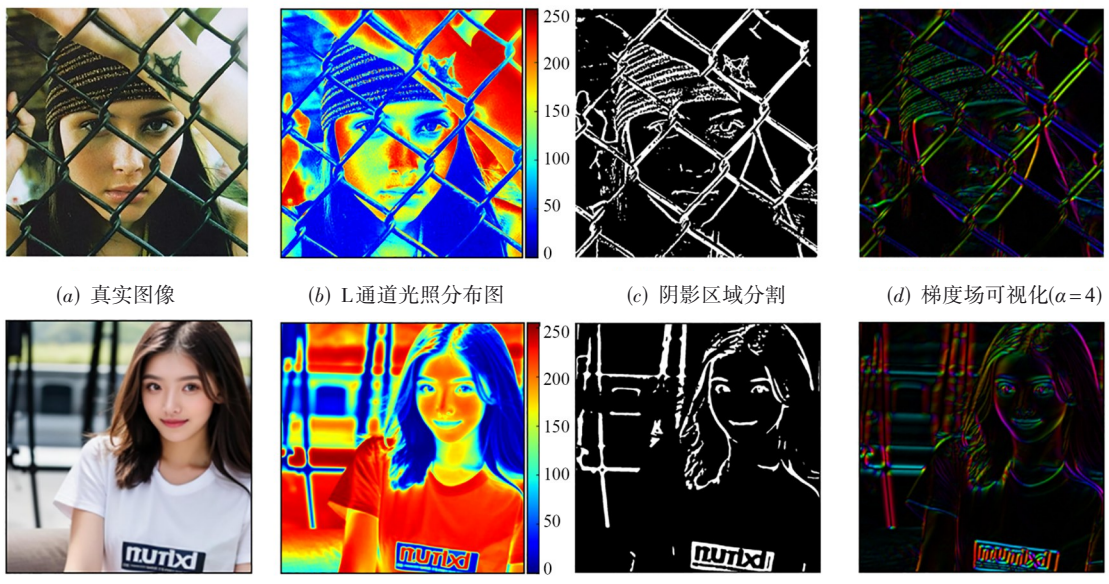


图2 L-KAN整体框架图



(a) 真实图像 (b) L通道光照分布图 (c) 阴影区域分割 (d) 梯度场可视化($\alpha=4$)

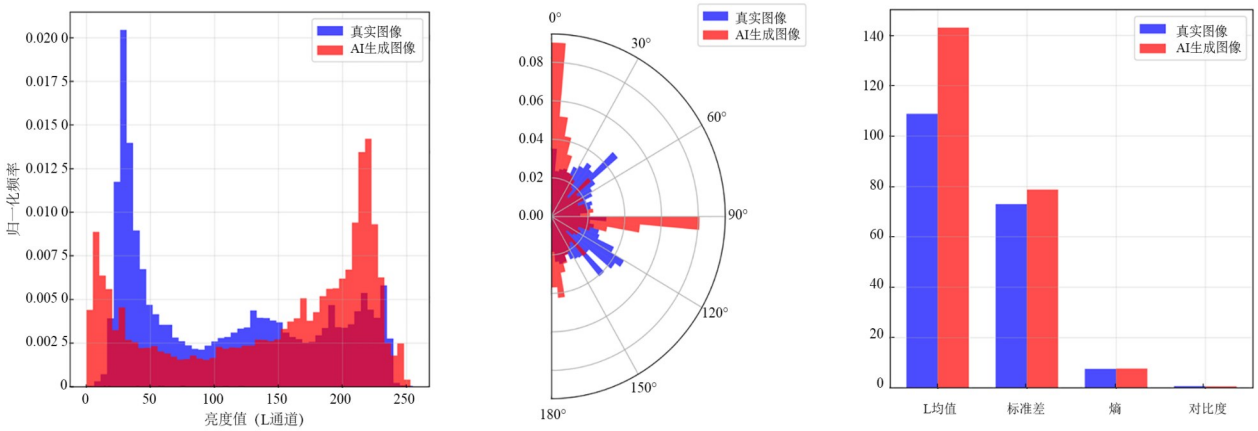
(e) AI图像 (f) L通道亮度分布 (g) 阴影区域分割 (h) 梯度场可视化($\alpha=4$)

图3 真实与生成图像多维分析对比图

真实与AI生成图像的多维特征差异分析表明生成模型在物理光照模拟方面存在明显局限. L通道亮度分布图中,真实图像呈现自然光谱过渡与不规则光照遮挡分布,而AI生成图像则表现为过度规则的环状面部轮廓、机械化的彩虹色带渐变和非自然垂直边缘切割,材质间热力特征差异不足. 阴影区域分割图显示,真实图像阴影具有复杂交叉遮挡与自然明暗过渡,AI生成图像则呈现简化的轮廓与边缘处理.

梯度场可视化($\alpha=4$)进一步揭示,真实图像保持主体边缘连续渐变,AI生成图像则存在局部细节突变与梯度场紊乱.

为定量分析真实图像与AI生成图像在光照分布上的差异,本文采用统计学方法对两类图像进行对比分析,结果如图4所示.



(a) 亮度直方图对比　　(b) 阴影方向分布对比　　(c) 统计量对比

图4　真实与生成图像多维量化分析对比图

亮度分布直方图显示,真实图像(蓝色)主要集中在低亮度区域(0~50),呈现明显低亮度峰值;而AI生成图像(红色)则集中在高亮度区域(175~225),形成显著高亮度集中,反映AI倾向生成较亮画面. 阴影方向分布玫瑰图中,真实图像方向分布较为均匀且数值小;而AI生成图像在约90°方向有明显峰值,表明存在方向性偏好,缺乏自然方向多样性. 统计量对比柱状图显示,AI生成图像在L均值、标准差上均明显高于真实图像,熵和对比度接近于真实图像. 这些特征差异揭示了AI生成图像在亮度分布、方向特性及统计特征上与真实图像的存在明显差异.

基于上述分析,本文创新性地引入光影敏感特征,从以下三个方面对光影敏感特征进行量化表征:

$$\boldsymbol{F}=\left[\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{s}},\boldsymbol{f}_{\mathrm{m}}\right] \tag{2}$$

该特征向量包含三类关键指标:

(1)基础光照统计量$\boldsymbol{f}_{\mathrm{b}}=\left(\mu_{\mathrm{Y}},\sigma_{\mathrm{Y}},H_{\mathrm{Y}}\right)$. 其中,$\mu_{\mathrm{Y}}$为亮度均值;$\sigma_{\mathrm{Y}}$为亮度标准差;$H_{\mathrm{Y}}$为亮度熵值,通过YUV色彩空间的亮度通道刻画全局光照分布.

(2)阴影几何特征$\boldsymbol{f}_{\mathrm{s}}=(R_s,\theta,\sigma_\theta)$. 其中,$R_s$为阴影面积比;$\theta$为主方向;$\sigma_\theta$为其标准差,基于自适应阈值分割提取阴影区域,反映光源与场景的物理一致性.

(3)多尺度光照梯度特征$\boldsymbol{f}_m=\left\{\left(U_s,M_s,\sigma_s,\bar{\theta}_s,\sigma_{\theta s},\rho_s\right)\right\}_{s\in\{1,2,4\}}$. 其中,$U_s$为光照均匀性;$M_s$、$\sigma_s$分别表示尺度$s$下的梯度幅值均值和标准差;$\bar{\theta}_s$、$\sigma_{\theta s}$分别表示梯度方向均值和标准差;$\rho_s$为其一致性,在三个空间尺度上分析光照均匀性、梯度幅值与方向统计量及其空间相关性,捕捉跨尺度光照异常.

3.2　KAF

在AI生成图像鉴别任务中,单一特征表示存在明显局限性,难以全面捕获生成图像与真实图像间的细微差异. 为解决此问题,本文综合利用RGB语义特征、频域特征、边缘特征和光影敏感特征构建多维表征体系. 然而,现有特征融合方法面临两大关键挑战:(1)融合过程中的噪声干扰问题,导致融合特征判别能力下降;(2)特征冗余与互补性的平衡困境,即难以在保留各特征流独特贡献时最小化冗余信息.

K-A表示定理严格证明:任意多元连续函数可分解为内外两层嵌套形式,通过有限次单变量非线性映射的层级组合即可逼近任意连续函数. 这一结果从数学上揭示了复杂映射的"可分解性",为特征融合提供了理论基础. 受此启发,本文将特征融合任务解构为内层非线性变换与外层非线性融合的协同过程,提出了基于K-A表示定理的特征融合框架(KAF),实现将复杂特征融合过程分解为两步简单且高效的函数变换.

在现有方法中,通常采用B样条函数构建K-A表示定理. 然而,初期实验表明,由于B样条基函数计算复杂度较高,训练效率显著下降. 径向基函数、傅里叶基函数和分段线性基函数等替代方案具有良好的理论性质,但在实际训练中仍面临显著的时间开销. 为优化计算效率,我们尝试通过批量化处理分段线性基函数,并引入低秩矩阵分解来降低参数规模,但效果仍未达到预期.

受Agarwal等人[36]提出的神经加性模型(Neural Additive Models,NAMs)启发,本文研究发现可以使用

基于ReLU激活函数的多层感知机(MLP)代替传统基函数. 具体而言,KAF由内层特征变换函数(Ψ函数)与外层特征融合函数(Φ函数)构成. Ψ函数通过线性变换、层归一化及非线性激活单元实现特征的深层表征;Φ函数则通过层次化神经网络架构完成多特征的融合与维度重构. 该设计显著提高了模型的计算效率和性能.

在上述KAF的基础上,本文采用级联方式构建双层KAF深度融合架构. 该架构将四路特征(RGB语义特征、频域特征、边缘特征及光影敏感特征)输入第一层KAF处理,随后将获得的融合特征替代原始RGB特征,与其余三种辅助特征共同输入第二层KAF进行深度融合. 这种特征替代机制充分利用了互补信息,有效增强了主干特征的表征能力,最终通过双层KAF级联结构提高了模型的鲁棒性与判别精度,如图5所示.

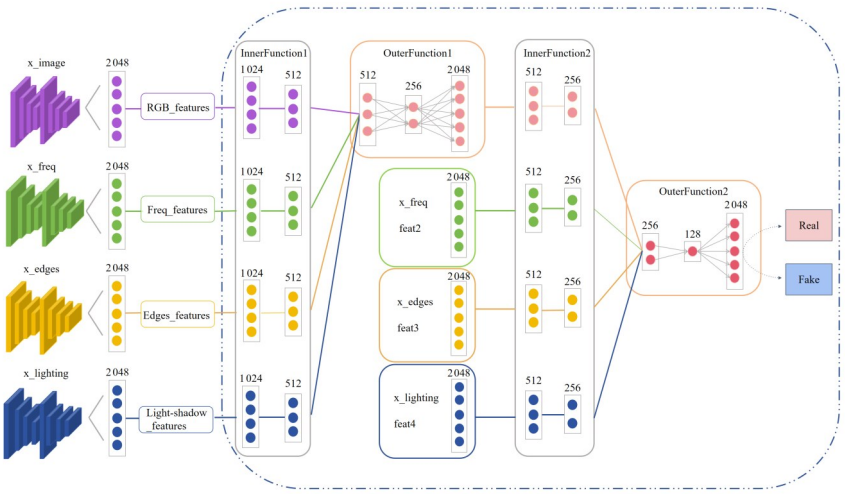


图5 级联KAF模块图

具体而言,对于给定的多特征$(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4)$,KAF的数学表示为

$$F_{\mathrm{KAF}}\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4\right)=\boldsymbol{\Phi}\left(\frac{1}{4}\sum_{i=1}^{4}\psi_i\left(\boldsymbol{x}_i\right)\right) \tag{3}$$

其中,ψ_i为内层函数;$\boldsymbol{\Phi}$为外层函数. 在内层非线性变换阶段,为确保各特征的独立性与通道正交性,对四路特征分别应用独立的内层函数ψ_i进行深度特征精炼,其中,内层函数ψ_i通过多层非线性映射捕获每一支路特征的判别性表征,非线性变换过程可简化公式形式表述如下:

$$\Psi_{p,q}(x)=\sum_{i=1}^{n}w_{p,q,i}\cdot\xi_i(x) \tag{4}$$

$$\xi_i(x)=\mathrm{ReLU}\left(\mathrm{LN}\left(\boldsymbol{W}_2\mathrm{ReLU}\left(\mathrm{LN}(\boldsymbol{W}_1x+\boldsymbol{b}_1)\right)+\boldsymbol{b}_1\right)_2\right) \tag{5}$$

其中,$w_{p,q,i}$为权重系数,p和q用于标识不同的特征分支,i表示对非线性变换结果$\xi_i(x)$的求和索引;$\boldsymbol{W}_1$为权重矩阵;$\boldsymbol{b}_1$为偏置向量增强模型表达能力;$\boldsymbol{W}_2$将第一次非线性激活后的输出映射到目标维度. 通过此过程,有效抑制了各特征支路间的相互干扰,确保了特征变换的独立性与判别性.

在外层特征融合阶段,采用计算高效且表达能力强的融合策略,实现对多特征的高阶语义整合,且通过非线性映射实现特征的深度融合与提纯:

$$\boldsymbol{\Phi}_q(y)=\sum_{j=1}^{m}v_{q,j}\cdot\xi_j(y) \tag{6}$$

$$\xi_j(y)=\mathrm{ReLU}\left(\mathrm{LN}\left(\boldsymbol{V}_1y+c_1\right)_j\right) \tag{7}$$

其中,q标识不同的融合分支;j对应基函数$\xi_j(y)$的索引;$v_{q,j}$决定最终融合特征的权重;$\boldsymbol{V}_1$为内层权重矩阵;c_1作为偏置向量增强模型对特征分布的适应性;LN表示层归一化,确保基函数$\xi_j(y)$的数值稳定性. 通过这个过程,有效捕获多特征间的高阶相关性,实现对融合特征的精准拟合与噪声抑制.

基于此结构,本文进一步设计了级联KAF架构,通过多层次特征融合提升模型表达能力. RGB语义特征通过ResNet50提取,包含丰富的语义信息和视觉内容,将其设置为主干特征,提供强大的表达能力. 频域特征、边缘特征和光影敏感特征则作为辅助特征,提供互补信息. 在网络实现中,构建了两个级联的结构,分别使用不同的维度设置,差异化设计既保证了初始特征的充分表达,又确保了最终融合特征的紧凑性与判别力.

在前向传播过程中,主干RGB语义特征在每个KAF模块中被递进更新,而辅助特征保持不变:

$$\boldsymbol{x}_1^{(l+1)}=F_{\mathrm{KAF}}^{(l)}\left(\boldsymbol{x}_1^{(l)},\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4\right) \tag{8}$$

其中,$\boldsymbol{x}_1^{(l)}$表示第l层的RGB语义特征;$\boldsymbol{x}_1^{(l+1)}$表示第$l+1$层更新后的特征;$\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$分别表示保持不变的频域特

征、边缘特征和光影敏感特征;$F_{KAF}^{(l)}$表示第l层KAF模块的特征融合函数.

通过这种递进式更新,主干特征能够在每一层逐步融合辅助特征提供的互补信息,同时保持辅助特征的原始判别性. 这种设计不仅建立了清晰的特征层次关系,还有效缓解了深度融合网络中的梯度消失问题和特征冗余. 通过理论驱动的设计,KAF实现了高效的异构特征融合,为AI生成图像识别提供了强有力的特征融合框架.

3.3　分类器设计

分类器的设计对系统整体性能起着决定性作用. 本文提出一种基于双层全连接神经网络的分类器结构,该结构在计算复杂度和特征表达能力之间实现了有效权衡. 具体而言,第一层网络采用线性变换将2 048维融合特征降维至512维特征空间,随后经过层归一化处理消除特征分布偏移,并通过ReLU非线性激活函数增强模型的表达能力. 为抑制过拟合现象,在第一层输出端引入dropout正则化机制,dropout率设为0.6. 第二层网络将512维特征映射至二维类别空间,最终经由Softmax函数得到样本类别概率分布.

实验结果表明,所提出的分类器结构在多组对照实验中表现出良好的性能. 其中,512维的隐层特征空间既保持了较强的特征表达能力,又有效控制了模型参数规模;0.6的dropout率能够显著抑制过拟合,提升模型在测试集上的泛化性能. 值得注意的是,层归一化的引入不仅加快了模型收敛速度,还增强了分类器对不同数据分布的适应能力. 实验表明,该分类器虽结构简洁,但具有充分的模型容量,能够有效刻画真假图像的判别边界.

3.4　损失函数

损失函数用于衡量模型预测与真实标签之间的差异,指导模型优化. 交叉熵损失函数(cross-entropy loss)通过最小化预测类别分布与真实标签分布的差异来优化模型,常与Softmax结合使用,从而提高分类准确性[37]. 另一方面,Li等人[38]在2021年提出单中心损失(Single-Center Loss,SCL),旨在通过仅压缩自然人脸的类内变异性同时增强类间差异,解决了传统Softmax损失无法显式促进类内紧凑性和类间可分性的问题. 该损失函数使神经网络能够在优化难度较低的情况下学习到更具判别力的特征表示. 本文采用了组合损失函数策略,结合交叉熵损失和单中心损失,以实现更有效的图像真伪鉴别. 总体损失函数表达式为

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{CE}}+\lambda\mathcal{L}_{\text{SC}} \tag{9}$$

其中,λ为权重系数,经实验验证设为0.1,用于平衡两种损失的贡献.

交叉熵损失$\mathcal{L}_{\text{CE}}$是分类任务的标准损失函数,数学表达式为

$$\mathcal{L}_{\text{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log p_i+\left(1-y_i\right)\log\left(1-p_i\right)\right] \tag{10}$$

其中,N表示批次大小;y_i是样本i的真实标签(0或1);p_i是模型预测样本i图像真实的概率.

单中心损失$\mathcal{L}_{SC}$的计算基于特征向量与自然人脸中心的距离度量,数学表达式为

$$\mathcal{L}_{\text{SC}}=M_{\text{nat}}+\max\left(M_{\text{nat}}-M_{\text{man}}+m\sqrt{D},0\right) \tag{11}$$

其中,M_{nat}和M_{man}分别为自然样本和篡改样本到中心点的欧式距离;$m\sqrt{D}$是距离阈值.

组合损失函数中,交叉熵损失确保模型在类别判别上的基础性能,而单中心损失通过优化特征空间分布提升模型的判别能力,二者协同作用显著提高了真伪图像鉴别准确率. 实验证明,当$\lambda=0.1$时,模型在保持训练稳定性的同时获得最佳性能表现.

4　实验结果与分析

4.1　数据集

为保证实验对比的一致性,本文采用ForenSynths训练集[18]对模型进行训练,训练集图像如图6所示. 该训练集由20个独立类别组成,每个类别包含约18 000张由ProGAN[39]生成的合成图像,以及相同数量的来自LSUN数据集[40]的真实图像. 参照现有研究方法,本文选取4类(car、cat、chair、horse)特定类别的数据来构建训练集.

(a) 真实图像　　(b) 生成图像

图6　ForenSynths训练集图像示例

为全面评估所提方法在真实应用场景下的泛化能力，本文构建三组测试集包含多种真实图像以及不同GAN和扩散模型生成图像的综合评估数据集，涵盖多种生成技术. 第一组测试集来自ForenSynths测试集[18]，其中AI生成图像由ProGAN[39]、GauGAN[41]、StarGAN[42]和Deepfake[43]生成，真实图像分别来自LSUN[40]、ImageNet[44]、CelebA[45]、Celeb-A-HQ[39]、COCO[46]和FaceForensics++[40]这6个公开数据集. 为模拟开放场景中的不可预测性. 第二组测试集来自GANGen9 GANs[21]数据集，其中AI生成图像由CramerGAN[47]、M-MDGAN[48]、RelGAN[49]和STGAN[50]等生成模型产生，真实图像从公开数据集(如LSUN[40]、ImageNet[44]、CelebA[45]、COCO[46])中采样，用于评估算法在未见生成模型上的检测能力. 实验数据配置如表1所示. 第三组测试集为来自NPR[21]提供的4种扩散模型数据集，由Stable Diffusion v2[51]、Midjourney[52]、VQ-Diffusion[53]和Glide[54]等扩散模型生成.

表1 实验数据设置

数据集	名称	图像尺寸	正样本数	负样本数	数据来源
训练集	train	(256, 256)	72 012	72 012	FS-Train (StyleGAN2-ADA)
第一组测试集	ProGan	(256, 256)	800	800	FS-Test (ProGAN)
	GauGAN	(256, 256)	5 000	5 000	FS-Test (GauGAN)
	StarGAN	(256~400, 256~400)	1 999	1 999	FS-Test (StarGAN)
	Deepfake	(256~400, 256~400)	2 707	2 707	FS-Test (DeepFake)
第二组测试集	CramerGAN	(128, 128)	2 000	2 000	GANGen (CramerGAN)
	MMDGAN	(128, 128)	2 000	2 000	GANGen (MMDGAN)
	RelGAN	(128, 128)	2 000	2 000	GANGen (RelGAN)
	STGAN	(128, 128)	2 000	2 000	GANGen (STGAN)
第三组测试集	Stable Diffusion v2	(256~1 102, 256~457)	1 000	1 000	NPR
	Midjourney	(256, 256)	2 000	2 000	NPR
	VQ-Diffusion	(256~1 102, 256~457)	1 000	1 000	NPR
	Glide_50_27	(256, 256)	1 000	1 000	NPR

4.2 实验结果

4.2.1 数据集分析

为验证光影敏感特征的有效性，本文对STGAN数据集进行分析，并通过特征分布可视化方法呈现结果. 如图7所示，真实图像与生成图像在光影敏感特征维度上表现出显著的统计差异.

图7的特征分布揭示了真实与伪造图像的显著差异. 在基础光照统计特征中，真实图像的Y通道亮度分布呈现较低的概率密度且分布自然分散，而伪造图像在50~150区间呈现异常的高概率密度集中；且AI生成图像的亮度标准差和熵均呈现异常的高峰值. 阴影特征方面，真实图像的阴影比例分布呈现自然离散特征，而AI生成图像表现出显著的高密度集中性；阴影角度分布上，两类图像均呈双峰值分布，但峰值仍存在细微差别；而在阴影角度标准差上，AI生成图像在低角度区域呈现出低于自然水平的分布密度，而高角度区域则出现显著的密度集中现象，反映了其在阴影细节特征上的统计异常性，违背了自然图像的基本规律. 多尺度光照梯度特征分析显示，三个尺度上AI生成图像均呈现出显著高于真实图像的概率密度集中性，反映了其在多层次光照细节上的统计特性偏差. 这些实验结果证实了所提光影敏感特征能够有效捕捉AI生成图像与真实图像的本质差异，为高精度伪造图像检测提供可靠的依据.

4.2.2 对比实验

为全面评估所提方法在生成图像鉴别任务的性能，将其与9种先进的图像伪造检测算法进行对比：CNNDetection[18]、Frank[22]、F3-Net[10]、SelfBland[55]、GANDetection[6]、LGrad[19]、Ojha[20]、NPR[21]和Rine[4]. 这些方法代表不同技术路线：基于像素空间特征学习(CNNDetection)、基于频率域分析(F3-Net、Frank)、基于梯度表示(LGrad)、基于知识注入与多任务学习(SelfBland)、面向跨模型泛化(Ojha、GAND-etection)、基于上采样伪影分析(NPR)以及基于中间层特征提取(Rine). 实验中，本文采用CNNDetection、Frank、F3-Net、LGrad、Ojha和NPR的官方开源代码进行复现，各算法参数配置均参照原始文献设定；Rine直接使用其最佳预训练模型；SelfBland、GANDetection使用文献[21]中的数据. 实验采用准确率(ACCuracy，ACC)和平均精度(Average Precision，AP)进行综合评估，其中ACC反映模型整体判别能力，AP通过计算PR曲线下的面积来评估模型在不同阈值下的综合表现.

在第一组数据集和第二组数据集上与前文所述9种方法进行对比的实验结果如表2和表3所示. 实验结果(表2和表3)表明，所提方法在两组测试集上均实现了优异的检测性能. 在第一组测试集上，本方法的

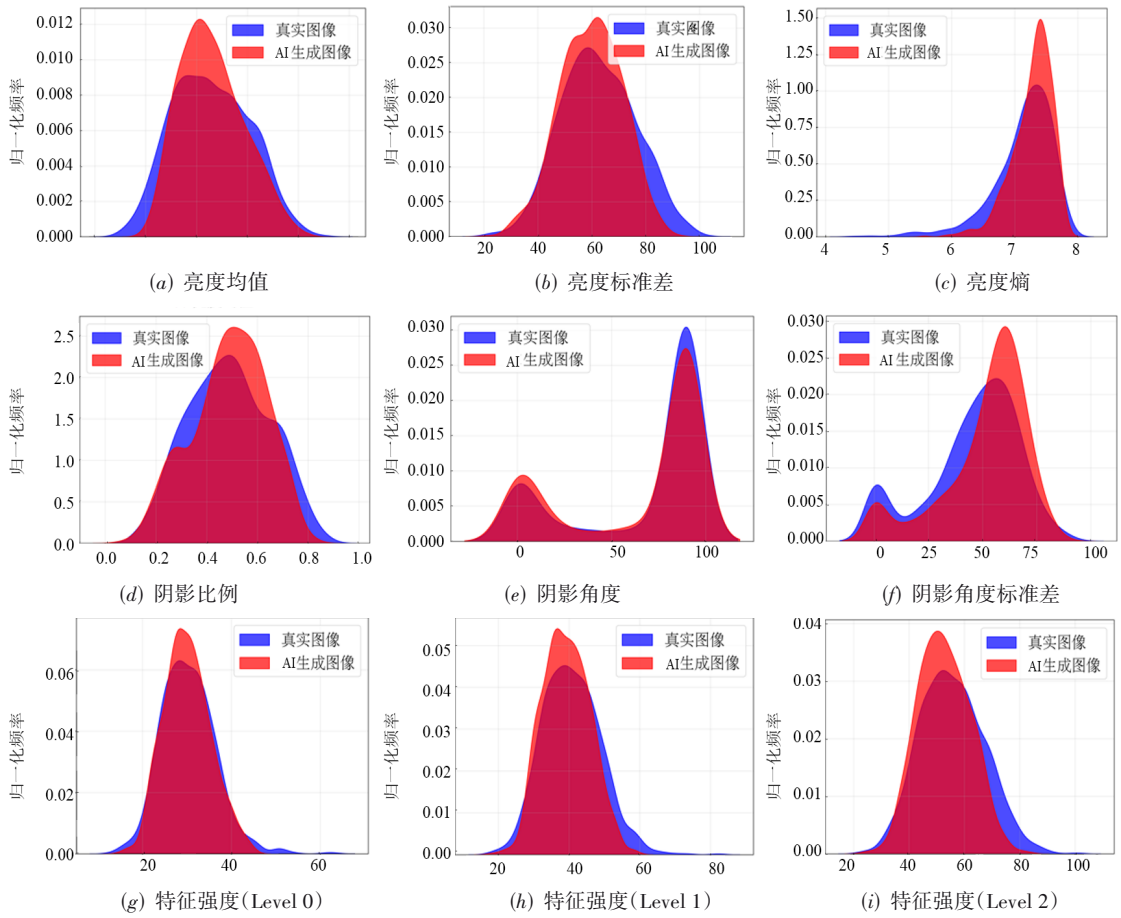


图7　光影特征分析图(以STGan为例)

表2　各模型在第一组数据集上的性能对比　　单位:%

模型	年份(会议/期刊)	ProGAN		GauGAN		StarGAN		Deepfake		mean	
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP
CNNDetection	2020 CVPR	99.3	100.0	57.1	62.5	95.1	100.0	63.8	91.6	78.8	88.5
Frank	2020 PMLR	90.3	85.2	68.8	74.8	98.8	98.8	60.7	49.1	79.7	77.0
F3-Net	2020 ECCV	99.7	99.7	56.9	62.1	100.0	100.0	71.5	87.7	82.0	87.4
SelfBland	2022 CVPR	58.8	65.2	59.2	65.5	74.5	89.2	93.8	99.3	71.6	79.8
GANDetection	2022 ICIP	82.7	95.1	61.4	75.8	68.8	99.7	60.0	83.9	68.2	88.6
Ojha	2023 CVPR	95.8	99.4	93.3	98.5	81.0	96.7	73.0	82.1	85.8	94.2
LGrad	2023 CVPR	99.9	100.0	56.9	80.4	100.0	100.0	66.1	83.9	80.7	91.1
NPR	2024 CVPR	99.8	100.0	85.8	87.7	99.7	99.0	77.4	86.2	90.7	93.2
Rine	2024 ECCV	99.8	100.0	99.7	100.0	96.4	100.0	60.8	97.0	89.2	99.2
Ours	2025	99.8	100.0	85.1	92.9	99.5	100.0	88.3	86.8	93.2	94.9

平均分类准确率和平均精度分别达到93.2%和94.9%.相比NPR方法(准确率90.7%,精度93.2%),平均分类准确率和平均精度分别提升2.5个百分点和1.7个百分点;相比Rine方法(准确率89.2%,精度99.2%),分类准确率提升4.0个百分点,而平均精度降低4.3个百分点.在第二组测试集上,本方法的平均分类准确率和平均精度分别达到96.4%和99.5%.相比NPR方法(准确率95.0%,精度98.9%),分类准确率和平均精度分别提升1.4个百分点和0.6个百分点;相比Rine方法(准确率93.4%,精度99.8%),平均分类准确率提升3.0个百分点,平均精度降低0.3个百分点.

细化分析表明,在第一组测试中,本文方法在最具挑战性的DeepFake样本上的检测性能(ACC 88.3%)相比其他方法有显著提升,较NPR的77.4%提高了

表3 各模型在第二组数据集上的性能对比 单位:%

模型	年份(会议/期刊)	CramerGAN		MMDGAN		RelGAN		STGAN		mean	
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP
CNNDetection	2020 CVPR	81.5	97.5	97.2	100.0	90.3	98.3	71.0	99.9	85.0	98.9
Frank	2020 PMLR	31.0	36.0	38.4	40.5	69.2	96.2	26.2	29.9	41.2	50.7
F3-Net	2020 ECCV	89.5	99.8	97.1	99.8	94.1	95.9	62.1	99.2	85.7	98.7
SelfBland	2022 CVPR	75.1	82.4	68.6	74.0	73.6	77.8	61.2	66.7	69.6	75.2
Ojha	2023 CVPR	77.6	99.8	79.4	97.2	77.3	89.9	72.6	90.4	76.7	94.3
LGrad	2023 CVPR	50.3	54.0	97.8	99.9	83.0	95.5	97.2	99.9	82.1	87.3
NPR	2024 CVPR	98.7	99.0	94.5	98.3	98.6	99.0	88.0	99.2	95.0	98.9
Rine	2024 ECCV	99.9	100.0	99.4	100.0	86.6	99.5	87.7	99.6	93.4	99.8
Ours	2025	98.2	100.0	98.2	100.0	93.6	98.1	95.7	99.9	96.4	99.5

10.9个百分点,较Rine的60.8%提高了27.5个百分点.在第二组测试中,本文方法在STGAN上取得了95.7%的ACC值,较NPR的88.0%提高了7.7个百分点,较Rine的87.7%提高了8.0个百分点.实验证明,本文方法在上述8个测试集上展现出优异的泛化能力和检测稳定性.

同时,为进一步测试所提方法在扩散模型上的性能,额外在第三组数据集上与CNNDetection[18]、Frank[22]、S-elfBland[55]、GANDetection[6]、Rine[4]这5种方法进行对比,实验结果如表4所示.

表4 各模型在扩散模型数据集上的性能对比 单位:%

模型	年份(会议/期刊)	Stable Diffusion v2		Midjourney		VQ-Diffusion		Glide_50_27		mean	
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP
CNNDetection	2020 CVPR	52.0	90.3	48.6	38.5	50.0	71.0	54.2	76.0	51.2	69.0
Frank	2020 PMLR	40.8	37.5	39.7	40.8	51.7	66.7	52.0	42.3	46.1	46.8
SelfBland	2022 CVPR	71.2	73.9	54.3	56.4	77.2	82.7	64.2	68.3	66.7	70.3
GANDetection	2022 ICIP	50.1	36.9	50.0	44.7	51.1	51.2	51.7	53.5	50.7	46.6
Rine	2024 ECCV	57.4	89.9	34.2	39.5	85.6	99.4	84.4	99.3	65.4	82.0
Ours	2025	66.3	99.5	51.8	56.4	80.2	99.9	74.2	85.1	68.1	85.2

实验结果表明,在平均准确率(ACC)上,本方法(68.1%)相比最好的SelfBland(66.7%)高1.4个百分点,且平均AP比SelfBland高出了14.9个百分点;在平均精度(AP)上,本方法达到85.2%,优于对比的其他方法;特别地,本方法在Stable Diffusion v2和VQ-Diffusion数据集上AP值接近100%. 值得注意的是,各方法在Midjourney数据集上检测效果普遍不佳,这揭示了当前AI生成图像检测技术面临的巨大挑战.

4.2.3 计算复杂度

计算复杂度是衡量算法实用性和部署可行性的关键指标. 本节对比所提出方法与现有主流生成图像检测方法在推理阶段的计算开销,详细结果见表5. 为确保公平比较,所有评估均基于统一的输入图像尺寸(224×224),并采用每次前向推理所需的浮点运算次数(Giga Floating-point Operations Per Second, GFLOPS)作为量化标准.

表5 各模型的计算复杂度对比

方法	CNNDetection	Frank	F3-Net	Self Bland	GANDetection	Ojha	LGrad	NPR	Rine	Ours
GFLOPS	10.4	10.5	32.6	8.8	129.3	10.4	5.4	3.5	52.0	24.6

结果显示,各方法的计算复杂度呈现较明显的分层特点. 低计算复杂度的方法包括NPR、LGrad和SelfBland. 其中,NPR的GFLOPS为3.5,这源于噪声残差预处理技术,通过捕获上下采样过程中的噪声实现伪造痕迹识别. 中等计算复杂度包括CNNDetection、Frank、Ojha和F3-Net. 其中,F3-Net(32.6 GFLOPS)因双分支结构和复杂频域变换导致计算量略高. 高计算复杂度包括Rine和GANDetection. 其中,G-ANDetection因特征图尺寸不缩减的设置(stride=1),使计算复杂度高至129.3 GFLOPS.

本文所提方法的GFLOPS为24.6,属于中等水平. 这主要是因为四路特征并行提取策略增加了一定的计算开销.

4.2.4 消融实验

为全面验证所提L-KAN模型各组件的有效性,本文设计了一系列消融实验,从基础模型逐步添加各核心模块,分析每个组件对整体性能的贡献. 表6展示了消

表6　各模块在第二组测试集上的消融实验结果　　单位:%

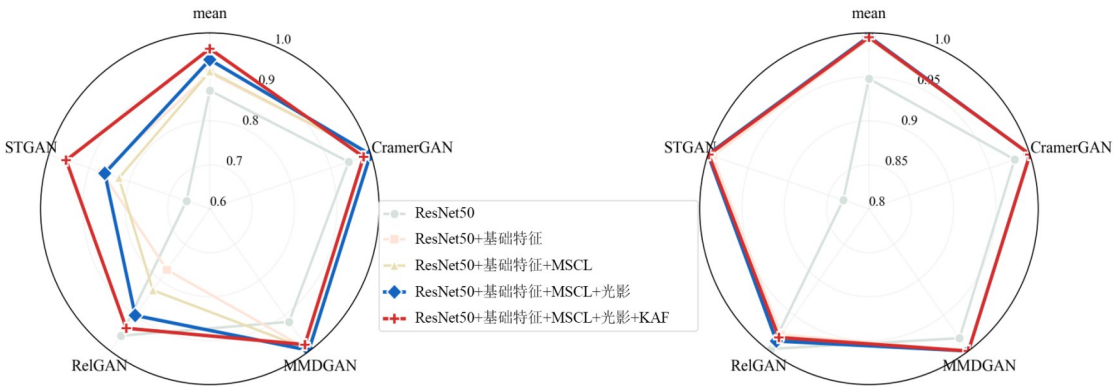
模块					CramerGAN		MMDGAN		RelGAN		STGAN		mean	
ResNet50	基础特征	光影	SCL	KAF	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP
√	×	×	×	×	94.5	98.1	91.8	98.2	95.7	99.7	65.8	83.2	86.9	94.8
√	√	×	×	×	99.8	100.0	99.7	100.0	77.2	97.7	86.8	99.4	90.9	99.3
√	√	×	√	×	99.9	100.0	99.5	100.0	82.9	98.6	82.6	99.8	91.2	99.6
√	√	√	√	×	99.9	100.0	99.8	100.0	90.0	98.6	86.1	100.0	93.9	99.6
√	√	√	√	√	98.2	100.0	98.2	100.0	93.6	98.1	95.7	99.9	96.4	99.5

融实验的详细结果,包括准确率(ACC)和平均精度(AP)两个指标.

在精心配置的数据增强和优化器等策略下,基础ResNet50网络已展现较高性能(平均ACC为86.9%,AP为94.8%).引入RGB、频域与边缘特征后,平均ACC提升至90.9%,AP提升至99.3%,证实多特征融合有效性.添加SCL模块使平均ACC微增至91.2%,在RelGAN数据集上ACC提升显著(5.7个百分点),增强了细微特征区分能力.随后,光影敏感特征的加入将平均ACC提高至93.9%,在RelGAN上表现突出(ACC提升7.1个百分点),验证了本文光照分析方法的关键作用.KAF模块的整合使平均ACC达到96.4%,在STGAN数据集上提升显著(ACC增长9.6个百分点),强化了模型泛化能力.完整L-KAN模型相较基准提升了9.5个百分点的ACC和4.7个百分点的AP,各组件呈现互补协同效应,显著增强了对多种GAN生成内容的检测鲁棒性.

为了更直观地展示光影特征与KAF模块对模型性能的贡献,本文绘制了如图8所示的雷达图,左侧为基于准确率(ACC)的性能对比,右侧为基于平均精度(AP)的性能对比,图中每条曲线代表不同配置的L-KAN模型在多种GAN生成器数据集上的表现.

从图8中可清晰观察到模型性能的渐进提升趋势.图8(*a*)ACC雷达图显示,基础ResNet50模型在CramerGAN和MMDGAN上表现较好,但STGAN上准确率仅为65.8%,存在明显短板.引入光影特征后,模型在RelGAN上性能显著提升,使五边形轮廓更加均衡;而KAF模块的加入使STGAN上的准确率大幅提高至95.7%,使整体轮廓近乎正五边形.图8(*b*)AP雷达图同样证实了完整模型在各数据集上均达到接近100%的高精度,验证了光影特征对光照不一致性的敏感捕捉能力和KAF模块对特征融合的增强作用,两者在提升模型泛化性和鲁棒性方面发挥了关键作用.



(*a*) ACC对比　　(*b*) AP对比

图8　L-KAN模型跨域检测性能雷达图

5　结束语

本文提出了一种融合光影敏感特征与K-A表示定理的生成图像检测方法(L-KAN).该方法通过分析基础光照统计特征(YUV色彩空间亮度分布)、阴影物理特征(区域分割与方向性分析)和多尺度光照梯度特征,有效识别生成内容在物理光学规律上的异常.基于K-A表示定理构建双层特征融合架构,解决了特征融合过程中的噪声干扰问题与特征冗余互补的平衡问题.实验结果表明,所提方法在两组测试集上均取得了优异性能,第一组测试集的平均分类准确率和平均精度分别达到93.2%和94.9%,第二组测试集分别达到96.4%和99.5%,第三组测试集分别达到68.1%和85.2%,相比现有最优方法在3组测试集上均实现了显著提升.未来研究将着重于光影敏感特征提取效率优化、特征融合机制改进以及跨媒体伪造检测扩展,为应对日益复杂的生成内容检测提供更完善的技术方案.

参考文献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[3] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[J].Advances Inneural Information Processing Systems, 2021, 34: 852-863.

[4] KOUTLIS C, PAPADOPOULOS S. Leveraging representations from intermediate encoder-blocks for synthetic image detection[M]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 394-411.

[5] WESTERLUND M. The emergence of deepfake technology: A review[J]. Technology Innovation Management Review, 2019, 9(11): 39-52.

[6] MANDELLI S, BONETTINI N, BESTAGINI P, et al. Detecting Gan-generated images by orthogonal training of multiple CNNs[C]//2022 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2022: 3091-3095.

[7] 朱世强, 王永恒. 基于人工智能的内容安全发展战略研究[J]. 中国工程科学, 2021, 23(3): 67-74.
ZHU S Q, WANG Y H. Development of content security based on artificial intelligence[J]. Strategic Study of CAE, 2021, 23(3): 67-74. (in Chinese)

[8] NIGHTINGALE S J, FARID H. AI-synthesized faces are indistinguishable from real faces and more trustworthy[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(8): e2120481119.

[9] 谢天, 于灵云, 罗常伟, 等. 深度人脸伪造与检测技术综述[J]. 清华大学学报(自然科学版), 2023, 63(9): 1350-1365.
XIE T, YU L Y, LUO C W, et al. Survey of deep face manipulation and fake detection[J]. Journal of Tsinghua University (Science and Technology), 2023, 63(9): 1350-1365. (in Chinese)

[10] ZHAO T C, XU X, XU M Z, et al. Learning self-consistency for deepfake detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 15003-15013.

[11] QIAN Y Y, YIN G J, SHENG L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[M]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 86-103.

[12] GRAGNANIELLO D, MARRA F, POGGI G, et al. Analysis of adversarial attacks against CNN-based image forgery detectors[C]//2018 26th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE, 2018: 967-971.

[13] VERDOLIVA L. Media forensics and DeepFakes: An overview[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 910-932.

[14] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1-9.

[15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[16] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[18] WANG S Y, WANG O, ZHANG R, et al. CNN-generated images are surprisingly easy to spot for now[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 8695-8704.

[19] TAN C, ZHAO Y, WEI S, et al. Learning on gradients: Generalized artifacts representation for gan-generated images detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 12105-12114.

[20] OJHA U, LI Y H, LEE Y J. Towards universal fake image detectors that generalize across generative models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 24480-24489.

[21] TAN C C, LIU H, ZHAO Y, et al. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 28130-28139.

[22] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//International Conference on Machine Learning. San Diego: PMLR, 2020: 3247-3258.

[23] TAN C C, ZHAO Y, WEI S K, et al. Frequency-aware

deepfake detection: Improving generalizability through frequency space learning[EB/OL]. (2024-05-12) [2025-06-05]. https://arxiv.org/abs/2403.07240v1.

[24] DURALL R, KEUPER M, KEUPER J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 7887-7896.

[25] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.

[26] 刘兵, 李穗, 刘明明, 等. 基于条件变分推断与内省对抗学习的多样化图像描述生成[J]. 电子学报, 2024, 52(7): 2219-2227.
LIU B, LI S, LIU M M, et al. Diverse image captioning via conditional variational inference and introspective adversarial learning[J]. Acta Electronica Sinica, 2024, 52(7): 2219-2227. (in Chinese)

[27] 张帅勇, 刘美琴, 姚超, 等. 分级特征反馈融合的深度图像超分辨率重建[J]. 自动化学报, 2022, 48(4): 992-1003.
ZHANG S Y, LIU M Q, YAO C, et al. Hierarchical feature feedback network for depth super-resolution reconstruction[J]. Acta Automatica Sinica, 2022, 48(4): 992-1003. (in Chinese)

[28] 林知心, 郑玉棒, 马天宇, 等. 基于轻量级全连接张量映射网络的高光谱图像分类方法[J]. 电子学报, 2024, 52(10): 3541-3551.
LIN Z X, ZHENG Y B, MA T Y, et al. Lightweight fully-connected tensorial mapping network for hyperspectral image classification[J]. Acta Electronica Sinica, 2024, 52(10): 3541-3551. (in Chinese)

[29] WILSON J, NICKISCH H, RÄTSCH G. Deep kernel learning[C]//International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 370-378.

[30] SYSKO-ROMAńCZUK S, STRISOVSZKY J, LUTSIV N. A representer theorem for deep kernel learning[J]. Joursnal of Machine Learning Research, 2022, 23(1): 1-32.

[31] 乔通, 陈彧星, 谢世闯, 等. 多色彩通道特征融合的GAN合成图像检测方法[J]. 电子学报, 2024, 52(3): 924-936.
QIAO T, CHEN Y X, XIE S C, et al. GAN synthetic image detection using fused features in the multi-color channels[J]. Acta Electronica Sinica, 2024, 52(3): 924-936. (in Chinese)

[32] KOLMOGOROV A N. On the Representation of Continuous Functions of Several Variables by Superpositions of Continuous Functions of a Smaller Number of Variables[M]. Providence: American Mathematical Society, 1961.

[33] SCHMIDT-HIEBER J. The Kolmogorov-Arnold representation theorem revisited[J]. Neural Networks, 2021, 137: 119-126.

[34] POLAR A, POLUEKTOV M. A deep machine learning algorithm for construction of the Kolmogorov-Arnold representation[J]. Engineering Applications of Artificial Intelligence, 2021, 99: 104137.

[35] LIU Z, WANG Y, VAIDYA S,et al. Kan: Kol-mogorov-arnold networks[EB/OL]. (2025-02-02) [2025-06-05]. https://arxiv.org/pdf/2408.02950.

[36] AGARWAL R, MELNICK L, FROSST N, et al. Neural additive models: Interpretable machine learning with neural nets[J]. Advances in Neural Information Processing Systems, 2021, 34: 4699-4711.

[37] GOODFELLOW I, BENGIO Y, COURSVILLE A, et al. Deep Learning[M]. Cambridge: MIT press, 2016.

[38] LI J M, XIE H T, LI J H, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6454-6463.

[39] KARRAS T, AILA T M, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. (2018-02-26) [2025-06-05]. https://arxiv.org/abs/1710.10196v3.

[40] YU F, ZHANG Y D, SONG S R, et al. Construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. (2016-06-04) [2025-06-05]. https://arxiv.org/abs/1506.03365.

[41] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 2337-2346.

[42] CHOI Y, CHOI M, KIM M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8789-8797.

[43] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++ : Learning to detect manipulated facial images[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1-11.

[44] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[45] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE,

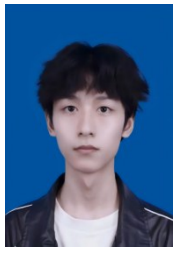
2015: 3730-3738.
[46] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
[47] BELLEMARE M G, DANIHELKA I, DABNEY W, et al. The Cramer distance as a solution to biased Wasserstein gradients[EB/OL]. (2017-05-30)[2025-06-05]. https://arxiv.org/abs/1705.10743v1.
[48] LI C L, CHANG W C, CHENG Y, et al. MMD GAN: Towards deeper understanding of moment matching network[EB/OL]. (2017-11-27)[2025-06-05]. https://arxiv.org/abs/1705.08584.
[49] NIE W L, NARODYTSKA N, PATEL A B. RelGAN: Relational generative adversarial networks for text generation[C]//International Conference on Learning Representations. Washington: ICLR, 2018: 1.
[50] LIU M, DING Y K, XIA M, et al. STGAN: A unified selective transfer network for arbitrary image attribute editing[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3668-3677.
[51] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10674-10685.
[52] RUSKOV M. Grimm in wonderland: Prompt engineering with midjourney to illustrate fairytales[EB/OL]. (2023-08-25)[2025-06-05]. https://arxiv.org/abs/2302.08961v2.
[53] GU S Y, CHEN D, BAO J M, et al. Vector quantized diffusion model for text-to-image synthesis[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10686-10696.
[54] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2022-03-28)[2025-06-05]. https://arxiv.org/abs/2112.10741v3.
[55] SHIOHARA K, YAMASAKI T. Detecting deepfakes with self-blended images[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18699-18708.

作者简介

邓 巧 女,2004年11月出生于湖南省衡阳市. 现为湖南工商大学人工智能与先进计算学院(湘江书院)本科生. 主要研究方向为计算机视觉.
E-mail: 1529776271@qq.com

姜 林 男,1977年11月出生于湖南省常德市. 现为湖南工商大学人工智能与先进计算学院(湘江书院)教授. 主要研究方向为智能语音处理、机器视觉、机器人应用.
E-mail: jlcdf@163.com

刘乐新 男,2004年1月出生于湖南省益阳市. 现为湖南工商大学人工智能与先进计算学院(湘江书院)本科生. 主要研究方向为计算机视觉.
E-mail: 207264603@qq.com



唐吕鑫 男,2005年2月出生于湖南省长沙市. 现为湖南工商大学人工智能与先进计算学院(湘江书院)本科生. 主要研究方向为计算机视觉.
E-mail: 3186221452@qq.com

杨英丽 女,2005年3月出生于江西省九江市. 现为湖南工商大学人工智能与先进计算学院(湘江书院)本科生. 主要研究方向为计算机视觉.
E-mail: 2307225478@qq.com