

融合色彩空间特征与频域特征的多源伪造图像检测方法

侯壹凡, 苗壮, 王俊, 张奇, 李阳, 王家宝

(陆军工程大学指挥控制工程学院, 江苏 南京 210000)

摘要: 随着人工智能技术的迅猛发展, 图像伪造手段变得日益复杂和隐蔽, 尤其是生成式伪造和图像编辑伪造手段, 使传统检测方法的泛化能力不足问题逐渐凸显。由于不同的伪造方式生成的图像存在显著差异, 现有研究往往局限于仅针对单一的伪造类型, 缺乏面向多源伪造的统一检测方法。为应对生成式伪造和图像编辑伪造带来的多源挑战, 提出了一种融合色彩空间特征与频域特征的多源伪造图像检测方法, 旨在提高当前图像伪造检测技术在复杂伪造场景下的泛化能力和鲁棒性。该方法从图像中提取色彩空间伪影特征、频域高频细节痕迹和语义信息, 通过构建的多域特征融合模块有效捕捉伪造图像中的光影异常与伪影, 实现对生成式伪造和编辑式伪造的多源伪造内容的准确识别。此外, 还构建了一个涵盖生成式伪造和编辑式伪造的多源伪造图像数据集, 更真实地模拟了现实中可能出现的伪造图像场景。实验结果表明, 所提方法在多源伪造图像检测任务中的性能显著优于现有方法, 尤其是在复杂伪造场景下表现出更强的适应性和准确性。

关键词: 多源伪造检测; 特征融合; 色彩空间; 小波特征; 注意力机制

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2026011

Multi-source forgery detection method based on the fusion of color space features and frequency domain features

Hou Yifan, Miao Zhuang, Wang Jun, Zhang Qi, Li Yang, Wang Jiabao

Institute of Command-and-Control Engineering, Army Engineering University of PLA, Nanjing 210000, China

Abstract: With the rapid development of artificial intelligence technology, image forgery techniques have become increasingly complex and covert, particularly in the areas of generative and image editing forgeries. Traditional detection methods have gradually revealed their limitations and shortcomings in dealing with these advanced forms of forgery. Due to significant differences in images generated by various types of forgeries, current research is often limited to detecting specific forgery types, with a lack of unified detection methods. To address the multi-source challenge posed by generative and image editing forgeries, a multi-source forgery image detection method that integrates color space features with frequency domain features was proposed. The proposed method aims to enhance the generalization ability and robustness of existing forgery detection techniques in complex forgery scenarios. This method extracts color space artifact features, frequency domain high-frequency detail traces and semantic information from the image. Through the constructed multi-domain feature fusion module, it effectively captures the abnormal light and shadow and artifacts in the forged image, achieving accurate identification of multi-source forged content including generative forgery and editing-based forgery. In addition, a multi-source forgery image dataset was

收稿日期: 2025-10-20; 修回日期: 2025-12-29

通信作者: 苗壮, emiao-beyond@163.com

constructed that includes both generative and editing forgeries, offering a more realistic representation of potential forgery scenarios in real-world applications. Experimental results demonstrate that the proposed method significantly outperforms existing methods in the detection of multi-source forgery images, especially in complex forgery environments, showing superior adaptability and accuracy.

Keywords: multi-source forgery detection, feature fusion, color space, wavelet features, attention mechanism

0 引言

图像伪造检测是一项借助人工智能技术对图像内容进行分析与判断,从而识别图像是否由人工智能生成或经过图像编辑工具篡改的技术^[1-2]。该技术主要依托深度学习、图像处理和模式识别等方法从图像的结构、纹理和色彩等维度挖掘潜在的伪造痕迹,实现对图像真伪的精准甄别^[3-4]。随着生成式人工智能的迅速发展,尤其是图像生成与编辑技术的飞速进步,伪造图像的手段变得日益复杂、隐蔽且难以辨识,极大地降低了传统检测方法的有效性。因此,开发具备鲁棒性的伪造图像检测技术来应对多源伪造图像,已成为信息对抗和维护国家信息安全的重要手段。

当前,伪造图像主要分为两类^[5],即生成式伪造图像和图像编辑伪造图像。生成式伪造图像^[6]通常借助生成对抗网络^[7]、扩散模型^[8]等深度生成技术从无到有生成,这类图像在内容、纹理和光照等方面与真实图像极为相似,极易“以假乱真”。图像编辑伪造图像^[9]主要通过 Photoshop (PS) 等图像处理工具对原始图像进行裁剪、拼接、重绘等操作实现,这类伪造图像基于真实图像,部分伪造痕迹(如边缘模糊、纹理断裂等)在高水平编辑下趋于消失。

尽管已有诸多研究针对特定伪造方式提出了有效的检测方法^[10-11],但现有方法大多仅面向单一的生成式伪造或图像编辑伪造^[12]检测,在面对多伪造源或未知伪造类型时,普遍存在鲁棒性差、泛化能力不足等问题^[5]。为更好地满足现实场景中多源、复杂的伪造图像检测需求,本文提出一种融合色彩空间特征、频域特征和语义信息的多源伪造图像检测方法。该方法不仅能够有效识别由生成式和图像编辑技术生成的伪造图像,还能够提升模型在面对多种伪造源时的鲁棒性和泛化能力。本文的主要贡献如下。

1) 提出一种融合色彩空间特征与频域特征的多源伪造图像检测框架,通过融合色彩空间特征、频域特征和语义信息,全面提取多源伪造图像中的共性特征,从而有效提高检测性能。

2) 设计了一个多域特征融合模块,通过注意力机制自适应地融合来自不同空间的特征,能够有效捕捉伪造图像中的伪影与光影特征,从而增强模型在生成式伪造图像与图像编辑伪造图像检测任务中的适应性与泛化能力。

3) 构建了一个多源伪造图像数据集 (multi-source forgery image dataset, MSFID), 该数据集涵盖了生成式伪造图像和多种图像编辑操作伪造的图像,为训练和评估具有更强泛化能力的检测模型提供了数据基础。在与基线模型的对比实验中,本文方法的检测效果显著优于现有方法,验证了其在多源伪造检测任务中的优越性。

1 相关工作

1.1 图像编辑检测

在图像编辑伪造领域,现有研究大多集中于特定类型的操作(如复制-移动、拼接和删除等)检测。早期的研究主要集中于利用手工或预制的特征(如频域特征^[13-14]、彩色滤波器阵列^[15]和离散余弦变换系数^[16])来检测伪造图像。随着深度学习技术的发展,研究者开始将深度学习技术应用于图像伪造检测。ManTra-Net^[17]通过提取图像的篡改特征,将图像篡改位置检测问题转换为局部异常检测问题;Kwon等^[18]构建了双流模型,通过分别从RGB流中学习视觉伪影、从离散余弦变换流中学习压缩伪影进行检测;Zhuo等^[19]设计了一个由粗到细的网络结构,结合空间和通道注意机制来进行伪造检测;Yu等^[20]提出一种结合宏观特征和微观特征的两阶段学习框架,通过多任务学习增强对篡改痕迹的捕捉能

力；此外，MVSS-Net^[21]通过多尺度监督、多视图特征学习，利用图像篡改边界伪影和输入图像的噪声视图进行检测。然而，上述方法通常局限于利用篡改的边界伪影的特征进行伪造检测，在面对生成式伪造图像时缺乏泛化性。本文方法同时针对图像编辑伪造与生成式伪造，综合提取多域特征，从而在多样化伪造场景下展现出较强的泛化能力。

1.2 生成式图像检测

Wang等^[22]提出了一种基于扩散重建误差的图像表示方法，通过比较输入图像与其通过预训练扩散模型重建的图像之间的误差，为针对扩散模型生成的图像的检测提供了一种的思路；Guo等^[5]利用伪造区域的分布特性与伪造方法的相关性，提出了一种分层细粒度伪造属性分类方法；Frank等^[23]从频域出发，分析了多种生成对抗网络模型生成的伪造图像，提出了基于频域特征的伪影检测分类方法；Liu等^[24]聚焦于生成对抗网络模型生成的人脸伪造图像，通过Gram模块提取人脸的纹理特征来增强检测器基于真伪图像在频率成分和频率统计上的差异进行人脸伪造检测的泛化性；检测方法F³-Net^[25]同样利用频域信息，基于真伪图像在频率成分和频率统计上的差异进行人脸伪造检测。然而，这些方法大多侧重于从生成过程中寻找伪造特征，通常仅适用于一种或少数几种生成模型，未能涵盖图像编辑类伪造图像的检测。本文所构建的数据集采用当前广泛使用的消融扩散模型（ablated diffusion model, ADM）生成生成式伪造图像，同时包含多种图像编辑伪造图像。通过融合提取色彩空间特征与频域特征，旨在捕捉不同伪造来源下的共性伪造痕迹，从而提升模型在多源伪造场景中的泛化检测能力。

2 方法

现有研究表明，融合空域与频域特征有助于更有效地揭示图像中的伪造痕迹。多源伪造图像检测的关键在于模型能否准确辨识不同伪造来源所遗留的独特特征模式。基于图像编辑的伪造图像通常在颜色分布和纹理渲染上存在不一致，而生成式伪造图像则容易在频域中引入特定的伪影模式。仅依靠单一图像域的特征难以全面捕捉多

样化的伪造痕迹，因此融合多域特征成为应对该挑战的有效技术途径。为此，本文提出一种融合色彩空间特征与频域特征的多源伪造检测框架，其核心设计思想如下。

1) 提取并增强HSV色彩空间中的信息，放大伪造区域与真实区域在颜色和光照方面的一致性。

2) 通过空域与频域并行的细节表征模块，捕捉图像在频谱维度上的异常特征。

3) 引入注意力机制，自适应地融合来自色彩特征、空域细节与频域细节的互补信息，使模型聚焦于最具判别力的特征。

该设计使模型能够同时从色彩一致性、频域统计异常和语义连贯性等角度挖掘伪造证据，形成更全面、互补的特征表示，从而实现对从传统编辑到现代人工智能生成等伪造图像的可靠检测。该框架的整体结构如图1所示。

2.1 色彩增强模块

在伪造图像检测任务中，图像的色彩信息往往蕴含重要的篡改线索，色彩异常通常被视为篡改操作的有效指示。为增强模型对色彩空间特征的提取能力，本文提出一种结合HSV通道加权与深度可分离卷积的色彩增强模块，旨在高效捕捉图像中的伪造痕迹。

色彩增强模块的设计思路如下：首先借助通道注意力机制为色调-饱和度-明度（HSV）空间的3个通道分配自适应权重，以强化对检测任务关键色彩特征的关注，提升色彩表征的判别性。具体实现中，先对HSV输入特征进行全局平均池化以压缩空间维度，再通过全连接层生成各通道的权重。这一过程的形式化表达如式(1)所示。

$$\begin{aligned} \text{Attention}_c(\text{hsv}) &= \sigma(\text{MLP}(\text{AvgPool}(X_{\text{hsv}}))) \\ \text{HSV}_{\text{feature}} &= X_{\text{hsv}} \otimes \text{Attention}_c(\text{hsv}) \end{aligned} \quad (1)$$

其中， X_{hsv} 表示图像在HSV色彩空间中的特征，AvgPool表示全局平均池化操作，MLP表示全连接网络， $\sigma()$ 为激活函数， $\text{Attention}_c(\text{hsv})$ 为注意力权重，最终通过权重对各通道进行加权，从而聚焦于关键色彩信息，在减少冗余特征的同时提升模型在色彩空间层面的感知能力。

同时，本文引入深度可分离卷积对加权后的特征进行高效处理。该结构将标准卷积分解为深

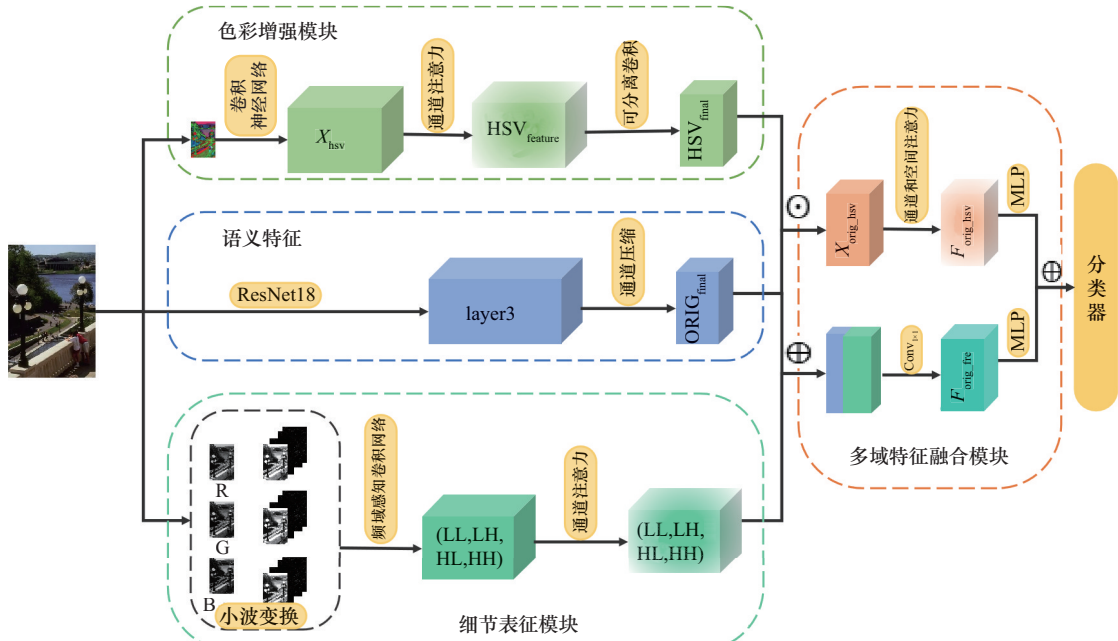


图1 融合色彩空间特征与频域特征的多源伪造图像检测模型的整体结构

度卷积和逐点卷积两步执行。具体而言，先利用深度卷积对加权后的各通道特征进行独立空间卷积，再通过 1×1 逐点卷积对多通道特征进行线性融合，得到最终输出特征 HSV_{final} 。该设计在保持较强特征提取能力的同时，显著降低了计算复杂度。此外，在每个卷积层后接入批归一化和高斯误差线性单元（GELU）激活函数，以增强网络的非线性表达能力。

2.2 细节表征模块

已有研究表明，图像伪造产生的伪影在频域中常呈现异常的频率分布，因而更易于被捕捉^[26-27]。受此启发，本文采用小波变换^[28]将图像从空间域转换到频域，以实现伪造痕迹的特征提取。对于每幅输入图像，本文分别对其各个通道进行二维离散小波变换，得到一个低频子带（LL）和3个高频子带（LH、HL、HH），变换过程如式(2)所示。

$$\begin{aligned}
 LL(i,j) &= \sum_M \sum_N I(m,n) \cdot \Phi(2i-m) \cdot \Phi(2j-n) \\
 LH(i,j) &= \sum_M \sum_N I(m,n) \cdot \Phi(2i-m) \cdot \Psi(2j-n) \\
 HL(i,j) &= \sum_M \sum_N I(m,n) \cdot \Psi(2i-m) \cdot \Phi(2j-n) \\
 HH(i,j) &= \sum_M \sum_N I(m,n) \cdot \Psi(2i-m) \cdot \Psi(2j-n)
 \end{aligned} \quad (2)$$

其中， M 和 N 分别表示图像的宽度与高度， $i = [0, \dots, M/2]$, $j = [0, \dots, N/2]$ 表示输出子带中的位置索引， $I(m,n)$ 表示输入图像在位置 (m,n) 处的像素值， Φ 和 Ψ 分别表示低通滤波器与高通滤波器。低频分量通常包含图像的结构信息，而高频分量则包含纹理等细节信息。

本文将3个高频分量（LH、HL、HH）进行堆叠，并对所有通道的振幅信息和相位信息分别进行整合，形成振幅特征与相位特征两组频域表示。随后，通过频域感知卷积网络（该网络由多个卷积层与批归一化层组成）对上述分解得到的振幅和相位图像特征进行提取与处理。为增强模型对异常频段的敏感性并提升对伪造痕迹的提取能力，在频域感知卷积网络处理之后，接着计算振幅和相位特征的各个通道的权重，通过加权融合得到最终的频域特征 FRE_{final} 。

2.3 多域特征融合模块

多域特征融合模块采用基于注意力的混合机制，将图像的色彩空间特征、频域特征与语义信息进行有效融合。本文通过色彩增强模块与细节表征模块分别提取色彩空间特征与频域特征，而语义信息则通过预训练模型获取。为平衡模型性能与计算效率，本文采用ResNet18^[29]作为特征提取骨干网络，并截取至layer3的输出作为语义信息 $ORIG_{final}$ 。

(1) 色彩空间特征与语义信息的融合

本文结合通道注意力和空间注意力机制，首先对 HSV_{final} 、 $ORIG_{final}$ 进行逐元素相乘操作，然后通过自适应平均池化进行压缩得到通道特征图；接着，使用两个卷积层依次进行通道数的压缩与恢复，其中第2个卷积层将通道数恢复到原始维度，并通过 Sigmoid 激活函数生成每个通道的权重 W ，该权重用于确定各通道的重要性；最后通过加权求和，得到通道融合特征。上述过程的形式化表示如下：

$$\begin{aligned} X_{orig_hsv} &= ORIG_{final} \odot HSV_{final} \\ W &= \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(\text{AvgPool}(X_{orig_hsv})))))) \\ F &= W \odot ORIG_{final} + (1 - W) \odot HSV_{final} \end{aligned} \quad (3)$$

其中， δ 表示 ReLU 激活函数， σ 表示 Sigmoid 激活函数， W 表示通道权重。

在 F 基础上，通过最大池化和平均池化分别提取不同的空间统计特征，将这两种统计特征拼接后，经由卷积层生成空间权重。随后，将通道加权融合后的特征 F 与该空间权重相乘，最终得到融合表征 F_{orig_hsv} 。这一过程的形式化表示如下：

$$\begin{aligned} W &= \sigma(\text{Conv}_{7 \times 7}([\text{AvgPool}(F), \text{MaxPool}(F)])) \\ F_{orig_hsv} &= W \otimes F \end{aligned} \quad (4)$$

其中， W 表示空间权重， F_{orig_hsv} 表示色彩空间特征与语义信息的融合特征。

(2) 频域特征与语义信息的融合

首先使用两个并行的卷积层将 FRE_{final} 、 $ORIG_{final}$ 的通道数减半，然后进行拼接，接着通过 1×1 卷积层得到融合后的特征 F_{orig_fre} 。这一过程的形式化表达如下：

$$\begin{aligned} \text{concat} &= [\text{Conv}(ORIG_{final}), \text{Conv}(FRE_{final})] \\ F_{orig_fre} &= \delta(\text{Conv}_{1 \times 1}(\text{concat})) \end{aligned} \quad (5)$$

最终，将 F_{orig_hsv} 与 F_{orig_fre} 进行拼接，得到特征向量 VECTOR，将其输入全连接网络进行分类（损失函数采用交叉熵）。

本文所提模型充分考虑了图像的色彩空间特征、频域特征和语义信息，通过多域特征融合模块结合注意力机制，有效整合了来自不同特征源的有效信息。总体而言，本文构建了一

个融合色彩空间特征与频域特征的统一范式：通过并行提取图像的色彩空间特征与频域特征，建立了一个覆盖更全面的伪造特征提取体系，并在此基础上设计了注意力引导的特征融合模块，该模块能够根据输入图像的内容动态调整各特征域的融合权重，而非采用固定的加权方式。这种自适应机制使得模型在面对不同类型的伪造图像时，能够优先选择最相关的特征进行决策，从而提升了对未知伪造技术的泛化能力。

3 实验

3.1 实验设置

(1) 数据集。本文采用的数据集信息见表1。本文构建的 MSFID 数据集包含 10 000 余幅伪造图像与 10 000 余幅真实图像。该数据集涵盖图像编辑（如拼接、复制-移动、删除）伪造图像和生成式伪造图像。MSFID 中的图像编辑伪造图像来源于 CASIA v2^[30-31]。CASIA v1 和 CASIA v2 专为图像伪造检测而设计，两者均是通过拼接、复制-移动等操作生成的伪造图像。MSFID 中的生成式伪造图像是使用在 ImageNet 数据集上预训练的开源 ADM^[32] 生成器，通过输入类别标签生成的，所生成的图像均经过人工抽样检查，从而确保样本质量。本文以 8:2 的比例将数据集划分为训练集和测试集。

表1 数据集信息

数据集	伪造类型	图像（真实/伪造）数量
CASIA v1	拼接、复制-移动、删除	800/921/幅
CASIA v2	拼接、复制-移动、删除	7 200/5 123/幅
MSFID	拼接、复制-移动、删除、生成式	10 000 余/10 000 余

(2) 评价指标。本文采用 F1 分数、ROC 曲线下面积（AUC）和准确率（ACC）3 项指标对模型性能进行综合评估。F1 分数可以衡量检测方法对伪造样本的识别能力与误判成本；AUC 通过计算 ROC 曲线下的面积来量化分类器

区分正负样本的能力。由于本文所构建的数据集中正负样本分布较均衡,使用准确率可以直观地反映检测方法的整体判别效果。

(3) 实验参数。本文模型基于 PyTorch 框架搭建,在 RTX 3090 GPU 环境下完成训练。实验设定批处理大小为 16,初始学习率为 0.001,并采用指数衰减策略进行动态调整。模型共训练 100 个 epoch,输入图像格式包括 JPEG、PNG 等常见格式。

3.2 对照方法

本文选取 F³-Net、MVSS-Net、PSCC-Net 和 TruFor 4 个具有代表性的检测方法作为基准,与本文方法进行对比实验。各基准模型的具体信息如下。

F³-Net^[25]: 该方法引入了频域感知分解 (frequency-aware decomposition, FAD) 与局部频率统计 (local frequency statistics, LFS) 两种频域特征提取模块,并设计了双路融合网络对两类特征进行融合,从而实现频域内图像伪造的检测。

MVSS-Net^[21]: 该方法采用多视图输入进行多尺度网络训练,能够有效提取图像的边界结构特征与整体语义信息。

PSCC-Net^[33]: 该方法提出了一种渐进式空间一通道网络,利用多尺度的特征以交叉连接方式提取粗粒度和细粒度的特征。

TruFor^[34]: 该方法融合 RGB 图像与噪声敏感指纹,基于 Transformer 架构同时提取高级与低级痕迹。

3.3 实验结果与分析

3.3.1 对比实验

表 2 展示了各方法在 MSFID 数据集上的实验结果,从中可以看出,本文所提模型在 MSFID 数据集上的表现优于其他基准模型。具体而言, MVSS-Net 在 3 个指标上的表现欠佳,原因在于其边缘监督分支主要关注篡改区域边缘的细微伪影特征,噪声敏感分支用于捕获篡改区域和真实区域之间的噪声不一致性,这种设计对传统图像编辑伪造较为有效,但在面对生成式伪造图像时,这两类分支的判别能力均被弱化,导致模型整体性能下降。PSCC-Net 的

性能相比 MVSS-Net 有所提升,这是因为该方法考虑了伪造区域的尺寸变化,通过融合局部与全局特征,从而能够适应多尺度伪造区域的检测需求; F³-Net 在所有基准模型中表现最为突出,这是因为其从频域角度出发,可以学到伪造图像的纹理特征及篡改边缘特征,能够较好地区分伪造与真实图像在频域中的差异。然而,与本文方法相比, F³-Net 在 3 个指标上仍存在一定差距。本文方法不仅融合了频域信息以有效提取伪造特征,还引入了色彩空间信息作为辅助决策依据,从而提升了检测性能。与 TruFor 相比,本文模型在各项指标上均具有明显优势。值得注意的是, TruFor 在 MSFID 数据集上的训练效果较差,甚至未达到 50.0% 的随机概率,这是因为 TruFor 所依赖的 Noiseprint++ 噪声指纹主要捕捉相机成像过程中的固有噪声模式,而扩散生成的图像噪声分布整体较为均匀,其与真实图像的差异更多体现在高频细节与纹理一致性层面,而非局部噪声异常,导致 Noiseprint++ 的判别能力下降,进而影响了 TruFor 的整体性能。

表 2 各方法在 MSFID 数据集上的实验结果对比

方法	F1 分数	ACC	AUC
F ³ -Net	83.7%	72.2%	88.7%
MVSS-Net	72.8%	61.3%	58.1%
PSCC-Net	82.9%	75.1%	82.2%
TruFor	49.4%	43.9%	35.8%
本文方法	90.4%	86.2%	92.2%

为评估模型的泛化能力,本文分别在 CASIA v1 图像编辑数据集和由 ADM 生成器新构建的生成式伪造图像数据集 (简称 ADM 数据集) 上进行了对比实验,结果见表 3,从中可以看出,本文方法在两个数据集上的泛化性能均显著优于其他基准模型。PSCC-Net 在生成式伪造图像数据集上展现出一定的泛化能力,但在 CASIA v1 数据集上表现一般;其他方法在两个数据集上的整体表现一般,其中 TruFor 在 ADM 数据集上的表现仍然较差。

表3 各方法在CASIA v1和ADM数据集上的实验结果对比

方法	CASIA v1 数据集			ADM 数据集		
	F1 分数	ACC	AUC	F1 分数	ACC	AUC
F ³ -Net	69.7%	53.6%	72.3%	66.7%	50.1%	93.7%
MVSS-Net	66.2%	56.1%	60.5%	65.6%	58.4%	64.4%
PSCC-Net	58.6%	51.4%	50.7%	74.8%	66.5%	96.0%
TruFor	58.2%	51.1%	50.9%	62.1%	47.1%	30.1%
本文方法	85.5%	85.6%	93.1%	80.7%	76.1%	94.7%

为直观地展示各方法的检测效果, 本文可视化了各方法在CASIA v1数据集和ADM数据集上的混淆矩阵, 分别如图2和图3所示。可以看出, MVSS-Net、PSCC-Net和F³-Net在两个数据集上多数样本被预测为负类, 而TruFor则倾向于预测为正类。这是因为ADM生成的伪造图像质量较高, 噪声分布较为均匀, 缺乏明显的篡改痕迹, 导致对照模型在特征提取过程中出现混淆, 未能有效学习到生成式伪造图像与编辑式伪造图像之间的区分性特征。

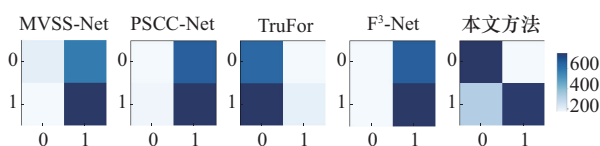


图2 各方法在CASIA v1数据集上的混淆矩阵

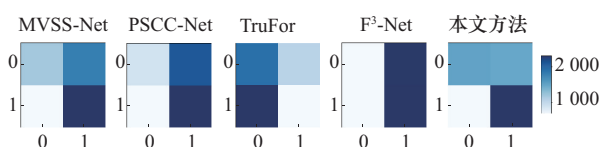


图3 各方法在ADM数据集上的混淆矩阵

3.3.2 各特征提取分支的重要性对比

本文针对色彩空间特征、频域特征和语义信息3个特征提取分支进行了重要性分析实验。实验中, 将 HSV_{final} 、 $ORIG_{final}$ 和 FRE_{final} 的初始权重均设为1, 训练完成后得到的结果如图4所示。从图中可以看出, 3个分支最终的权重系数均维持在1附近, 其中权重最高的分支相较于最低的分支仅高出0.03。这表明在训练过程中, 本文提出的模型能够均衡地学习3个分支的特征, 各分支的重要性基本一致, 验证了本

文提出的色彩增强模块对伪造图像检测具有有效贡献, 其重要性可以与骨干网络和频域特征提取部分相媲美。总体而言, 本文模型从色彩空间、频域特征和语义信息3个不同的角度提取图像特征, 3个分支对模型的贡献相对均衡。

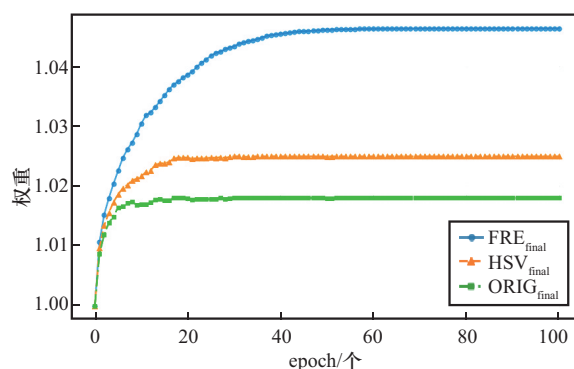


图4 各特征提取分支的重要性对比

3.3.3 收敛速度

本文对各基准模型的训练过程进行了可视化分析。各模型均训练100个epoch, 每4个epoch进行一次测试并记录结果, 提取了ACC与AUC两组指标的测试变化情况, 结果如图5和图6所示。可以看出, 本文模型在仅训练20个epoch时的性能已接近最佳性能(即训练100个epoch时的性能), 表现出显著的快速收敛能力。相比之下, 其他基准模型虽然在训练过程中性能逐步提升, 但在训练20个epoch时仍未达到本文模型的性能。这说明本文模型在训练效率和性能方面均具有优势, 不仅能够以更短的训练周期达到接近最优的性能, 在最终的测试中也展现出更优的准确性与收敛稳定性。

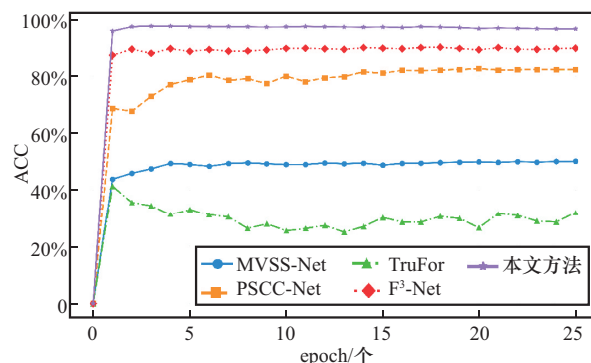


图5 训练过程中ACC指标趋势

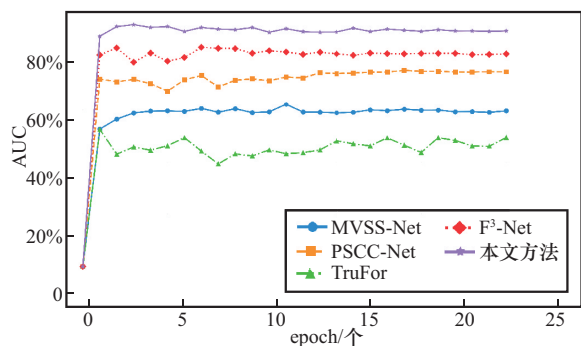


图6 训练过程中AUC指标趋势

3.3.4 可视化分析

为深入分析本文方法提取伪造样本特征的能力, 研究人员对 ResNet18 骨干网络 layer3 训练后的特征向量进行了可视化, 同时将本文方法最终提取的特征向量经过 t-SNE^[35]降维的结果也进行了可视化, 如图7和图8所示。可以看出, 本文方法提取的特征向量呈现较为清晰的类别边界, 表明本文方法能够在特征空间有效分离伪造样本与真实样本, 验证了其区分真伪的能力。

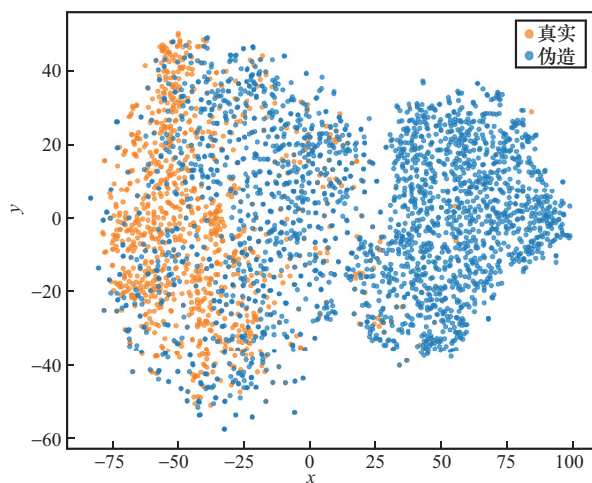


图7 ResNet18提取特征可视化散点图

此外, 从图7和图8中还可以观察到, 在伪造样本部分, 散点分布呈现较为分散的两部分, 而真实样本整体分布较集中。为进一步分析此现象, 本文对图像编辑和生成式两种伪造方式的特征向量分别进行了可视化, 结果如图9所示。可以看出, 伪造样本中的两个子类分别对应图像编辑伪造图像和生成式伪造图像, 表明本文模型不仅能有效检测伪造图像, 还能区分其伪造方式,

从而体现了模型在特征提取方面的能力。通过上述可视化分析, 验证了本文方法在提取真实与伪造图像高级特征方面的有效性, 并能够识别伪造图像的生成方式。

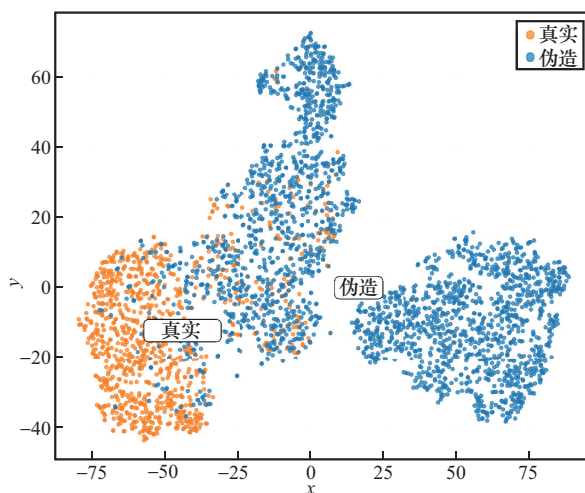


图8 本文方法提取特征可视化散点图

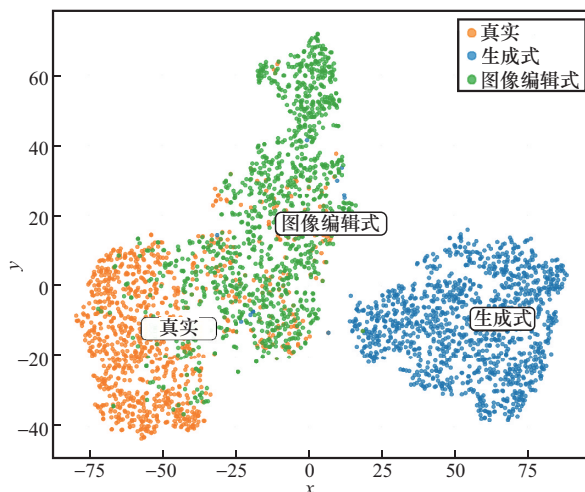


图9 本文方法提取特征区分伪造方式可视化散点图

4 结束语

本文针对多伪造源图像检测问题, 提出一种融合色彩空间与频域特征的多源伪造图像检测模型。通过结合色彩空间特征、频域特征及语义信息, 该模型能够有效识别基于生成式技术与图像编辑技术生成的伪造图像。实验结果表明, 与现有方法相比, 本文模型在多源伪造图像检测任务中表现出更强的鲁棒性和泛化能力。

然而, 本文构建的多源数据集 MSFID 中的生成式图像仅基于 ADM 生成器生成, 未来的研

究可以引入效果更好、更富挑战性的生成模型,构建更加多样化的伪造样本,以全面验证模型在多源场景下的泛化能力。此外,为应对日益复杂的实际应用需求,还需要探索具有更强表征能力的检测模型,从而提升对新型伪造技术的识别性能。

参考文献:

- [1] Zhai Y H, Luan T Y, Doermann D, et al. Towards generic image manipulation detection with weakly-supervised self-consistency learning[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2024: 22333-22343.
- [2] 黄诗瑀, 叶锋, 黄添强, 等. 人脸伪造与检测中的对抗攻防综述[J]. 网络与信息安全学报, 2023, 9(4): 1-15.
Huang S Y, Ye F, Huang T Q, et al. Survey on adversarial attacks and defense of face forgery and detection[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 1-15.
- [3] Zanardelli M, Guerrini F, Leonardi R, et al. Image forgery detection: a survey of recent deep-learning approaches[J]. Multimedia Tools and Applications, 2023, 82(12): 17521-17566.
- [4] 张博林, 朱春陶, 殷琪林, 等. 基于噪声注意力的伪造人脸检测方法[J]. 网络与信息安全学报, 2023, 9(4): 155-165.
Zhang B L, Zhu C T, Yin Q L, et al. Noise-attention-based forgery face detection method[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 155-165.
- [5] Guo X, Liu X H, Ren Z Y, et al. Hierarchical fine-grained image forgery detection and localization[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 3155-3165.
- [6] Zhu M J, Chen H T, Yan Q Y, et al. Genimage: a million-scale benchmark for detecting AI-generated image[C]//Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023(NeurIPS 2023). 2023.
- [7] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[C]//the 7th International Conference on Learning Representations (ICLR). 2019.
- [8] Ho J, Jain A, Abbeel P, et al. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 6840-6851.
- [9] Dagar D, Vishwakarma D K. A noise and edge extraction-based dual-branch method for shallowfake and deepfake localization[J]. Signal, Image and Video Processing, 2025, 19(3): 198.
- [10] 林善和. 基于孪生网络的鲁棒性深度伪造检测方法[J]. 网络与信息安全学报, 2024, 10(2): 182-189.
Lin S H. Robust deepfake detection method based on Siamese network[J]. Chinese Journal of Network and Information Security, 2024, 10(2): 182-189.
- [11] 王金伟, 张子荷, 罗向阳, 等. 基于强化样本的伪孪生网络图像篡改定位模型[J]. 网络与信息安全学报, 2024, 10(1): 33-47.
Wang J W, Zhang Z H, Luo X Y, et al. Pseudo-siamese network image tampering localization model based on reinforced samples[J]. Chinese Journal of Network and Information Security, 2024, 10(1): 33-47.
- [12] Du B, Jiang Z H, Lei Z Y, et al. IMDL-BenCo: a comprehensive benchmark and codebase for image manipulation detection & localization[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 134591-134613.
- [13] Li G H, Wu Q, Tu D, et al. A sorted neighborhood approach for detecting duplicated regions in image forgeries based on DWT and SVD[C]//Proceedings of the 2007 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2007: 1750-1753.
- [14] Zimba M, Xingming S. DWT-PCA (EVD) based copy-move image forgery detection[J]. International Journal of Digital Content Technology and Its Applications, 2011, 5(1): 251-258.
- [15] Popescu A C, Farid H. Exposing digital forgeries in color filter array interpolated images[J]. IEEE Transactions on Signal Processing, 2005, 53(10): 3948-3959.
- [16] Lin Z C, He J F, Tang X O, et al. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis[J]. Pattern Recognition, 2009, 42(11): 2492-2501.
- [17] Wu Y, AbdAlmageed W, Natarajan P. ManTra-Net: manipulation tracing network for detection and localization of image forgeries with anomalous features[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9535-9544.
- [18] Kwon M J, Yu I J, Nam S H, et al. CAT-Net: compression artifact tracing network for detection and localization of image splicing[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2021: 375-384.
- [19] Zhuo L, Tan S Q, Li B, et al. Self-adversarial training incorporating forgery attention for image forgery localization[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 819-834.
- [20] Yu Z Q, Ni J Q, Lin Y Z, et al. DiffForensics: leveraging diffusion prior to image forgery detection and localization[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 12765-12774.
- [21] Dong C B, Chen X R, Hu R H, et al. MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3539-3553.
- [22] Wang Z D, Bao J M, Zhou W G, et al. DIRE for diffusion-generated image detection[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:

IEEE Press, 2024: 22388-22398.

- [23] Frank J, Frank J, Eisenhofer T, et al. Leveraging frequency analysis for deep fake image recognition[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 3247-3258.
- [24] Liu Z Z, Qi X J, Torr P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8057-8066.
- [25] Qian Y Y, Yin G J, Sheng L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 86-103.
- [26] Durall R, Keuper M, Pfrendt F J, et al. Unmasking deepfakes with simple Features[J]. CoRR, 2020, 1911. 00686.
- [27] Yu N, Davis L, Fritz M. Attributing fake images to GANs: learning and analyzing GAN fingerprints[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7555-7565.
- [28] Finder S E, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 363-380.
- [29] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.
- [30] Dong J, Wang W, Tan T N. CASIA image tampering detection evaluation database[C]//Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing. Piscataway: IEEE Press, 2013: 422-426.
- [31] Pham N T, Lee J W, Kwon G R, et al. Hybrid image-retrieval method for image-splicing validation[J]. Symmetry, 2019, 11 (1): 83.
- [32] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 8780-8794.
- [33] Liu X H, Liu Y J, Chen J, et al. PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7505-7517.
- [34] Guillaro F, Cozzolino D, Sud A, et al. TruFor: leveraging all-round clues for trustworthy image forgery detection and localization[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 20606-20615.
- [35] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9: 2579-2605.

[作者简介]



侯壹凡 (1994-), 男, 河南鹤壁人, 陆军工程大学博士生, 主要研究方向为计算机视觉。



苗壮 (1976-), 男, 辽宁辽阳人, 陆军工程大学教授, 主要研究方向为图像视频处理。



王俊 (1997-), 男, 江苏盐城人, 陆军工程大学硕士生, 主要研究方向为智能安全。



张奇 (2000-), 男, 江苏徐州人, 陆军工程大学硕士生, 主要研究方向为图像复原。



李阳 (1984-), 男, 河北廊坊人, 陆军工程大学副教授, 主要研究方向为计算机视觉、图像检索



王家宝 (1985-), 男, 安徽合肥人, 陆军工程大学副教授, 主要研究方向为计算机视觉、图像处理。