

通过空间-频率协同学习和层次跨模态融合实现可泛化的深度伪造检测

Mengyu Qiao*
North China University of
Technology
Beijing, China
myuqiao@ncut.edu.cn

Runze Tian
North China University of
Technology
Beijing, China
trz@mail.ncut.edu.cn

Yang Wang
Ultramain Systems, Inc.
Albuquerque, NM, USA
ywang@ultramain.com

Abstract

深度生成模型的快速发展对深度伪造检测构成了关键挑战，因为训练在特定伪造特征上的检测器在遇到未见过的伪造时通常会出现显著的性能下降。虽然现有方法主要依赖于空间域分析，但频域操作主要限于特征级的增强，原生的频域伪造特征和空间频域交互却没有得到充分利用。为了应对这一限制，我们提出了一种新的检测框架，该框架集成了多尺度空间频率分析用于通用的深度伪造检测。我们的框架包含三个关键组成部分：(1) 一个局部光谱特征提取流程，结合块状离散余弦变换和级联多尺度卷积来捕捉微细的光谱伪造特征；(2) 一个全局光谱特征提取流程，使用尺度不变的差分累积来识别整体伪造分布模式；以及 (3) 一个多阶段的跨模态融合机制，结合浅层的注意力增强和深层的动态调制来建模空间频域交互。在广泛使用的基准测试上的深入评估表明，我们的方法在准确性和泛化性上均优于最新的深度伪造检测方法。

CCS Concepts

• Computing methodologies → Computer vision representations; • Information systems → Multimedia and multimodal retrieval.

Keywords

Deepfake Detection, Discrete Cosine Transform, Cross-Modal Fusion, Spatial-Frequency Learning, Convolutional Neural Network

1 引言

深度生成模型的快速发展 [?] [?] [?] 促进了面部操控技术的演变，使得高度逼真的合成媒体（通常称为“deepfakes”）的创建成为可能。尽管这些技术 [?] [?] [?] [?] 在娱乐和创意产业中具有潜在应用，但其滥用带来了显著的社会风险，包括虚假信息传播、身份盗窃和政治欺诈。因此，检测 deepfake 内容已成为多媒体安全和计算机视觉领域的一个关键挑战。

大多数现有的 deepfake 检测方法 [?] [?] [?] [?] 将该任务表述为二元分类，其中通常使用骨干网络来提取判别特征。分析和检测 [?] [?] [?] [?] [?] 主要关注空间域中的伪影，如面部不一致、不自然的纹理、混合边界等。然而，随着复杂的 deepfake 技术日益改进空间细节，像素级别上真实和伪造面孔之间的差异变得极其微妙且难以检测。这一局限性揭示了空间域方法的内在脆弱性，尤其是在面对高质量伪造或后处理扰动时。

相比之下，频域分析为揭示操控痕迹提供了一种补充视角。伪造的人脸通常表现出独特的频谱模式，例如高频噪声的不一致或生成模型中上采样操作引入的异常。尽管具有这种潜力，但现有的频率感知方法 [?] [?] [?] [?] [?] 通常通过在频

谱表示中的中间滤波操作增强空间异常，并在空间域中完成特征提取，忽略了跨频率子带中固有的异常模式，尤其是在局部伪造区域中，如图 ?? 所示。尽管有些工作 [?] [?] 结合了分块的频率变换，但它们未能深入研究频率特征之间的相互依赖关系。此外，大多数方法对图像语义与伪造伪迹之间的依赖性处理不足，导致在数据集和操作类型间的泛化能力有限。

为了解决这些挑战，我们提出了一种空间-频率协同学习 (SFCL) 框架，该框架能够同时从空间和频率域中提取特征，并通过分层跨模态融合 (HCMF) 机制实现多阶段特征交互。在频率域处理中，我们摒弃传统的全局频率变换，而采用局部化分析范式。具体而言，我们采用逐块离散余弦变换 (DCT) 来准确捕捉频域伪影，同时保留宏观层次的空间语义。为了建模块内 DCT 系数的分布模式和块间 DCT 子带的过渡模式，我们在频率管道中构建一个局部分支，通过跨频带和跨块卷积提取多尺度频率特征。此外，我们在频率管道中引入一个全局分支，从原始分辨率的人脸块的 DCT 子带的水平和垂直导数中计算统计频率特征。最后，我们引入了一个频率引导的注意力机制，以增强网络浅层的空间特征，同时在深层中实现有效的跨模态融合。

在基准数据集 (FaceForensics++, Celeb-DF 和 DFDC) 上进行的大量实验验证了我们方法在数据集内评估和跨数据集泛化中的卓越性能。总结而言，本研究的贡献如下三个方面：

- 我们提出了一种新颖的块状离散余弦变换分析方法，结合块内和块间多尺度频率卷积网络。通过在 DCT 变换域中的渐进特征抽象，此方法有效捕获块内的频谱特征和块间的过渡模式，从而增强了频域中的局部表示学习和精确伪影建模。
- 我们提出了一种基于尺度不变微分分析的全局频谱特征提取和合成方法。该方法模型化了感兴趣区域中的整体伪造模式，同时最小化了多尺度差异并减少了来自后期处理操作的干扰。
- 我们开发了一种层次化的跨模态融合机制，该机制整合了浅层注意力增强和深层动态调制。这个设计有效地建模了空间频率交互，解决了检测任务中的模态特异性特征异质性问题，并保持对复杂伪造操作的鲁棒性。

2 相关工作

早期检测方法集中在伪造中的非自然生物信号，例如讲话时异常的嘴巴运动 [?]、头部姿势 [?]、不规则的眨眼频率和远程视觉容积脉搏波图 (rPPG) 中的异常 [?] [?] [?] [?] [?]。

随着深度学习的进步，各种模型 [?] [?] [?] 作为主干网络被广泛应用于空间域中的深度伪造检测。为了进一步提升性能，许多方法 [?] [?] [?] [?] [?] 结合了专门的功能模块以增强对局部伪造模式的学习。例如，M2TR [?] 和 HFI-Net [?] 利用视觉变换器 (ViTs) [?] 和卷积神经网络 (CNNs) 的互补优势来

*Corresponding author.

提高检测准确性。MAD [?] 引入了多重注意力架构，配备有纹理增强块和区域独立损失，使网络能够关注不重叠的区域并捕捉细粒度的局部伪造线索。Nirkin 等 [?] 提出了一种基于面部区域与其周围上下文差异的方法，基于大多数操控技术仅修改面部而保留外围特征如头发、耳朵和颈部的观察。此外，一些方法通过将噪声特征 [?] [?] [?] [?] 与 RGB 特征集成来检测操控。

最近的研究 [?] [?] [?] [?] [?] [?] [?] 揭示了频域中的判别模式对于深度伪造检测的可喜效果。Qian 等人 [?] 提出了 F3-Net，通过 DCT 分离频率成分并融合局部频率统计，实现了压缩视频上的鲁棒性。Durall 等人 [?] 利用离散傅里叶变换 (DFT) 分析了高频谱异常，为频域伪造模式提供了理论见解。几种方法利用频域和 RGB 特征之间的互补性。Masi 等人 [?] 设计了一个双流网络，将高斯-拉普拉斯频率增强与空间特征相结合。Liu 等人 [?] 联合建模了利用相位谱和空间的上采样伪影。此外，空间频率关系 [?] [?] [?] [?] 的研究利用了多模态学习的表现能力。虽然这些方法利用了频率分析，但大多数仅仅依赖于基于滤波器的方法来增强空间域检测，缺乏对局部和全局尺度多样频率特征的深入探索，并忽视了不同阶段空间域和频域的交互作用。

3 方法

3.1 概述

本节介绍了我们提议框架的设计动机和架构概述。现代的深度伪造技术逐步从局部到全局合成内容，通常导致块间模式不一致，特别是在语义边界，例如面部区域。为了应对深度伪造伪影的多样模式，我们采用空间-频率协同学习范式，以实现全面表征和可推广的检测。我们提议框架的概述如图 1 所示。

在频域中，深度伪造操作往往在细粒度细节的合成过程中引入附加噪声或不自然的频谱扭曲。我们假设，应用逐块的 DCT 变换可以直接分析每个局部块中的高、中、低频伪影。通过将图像划分为多个区块，我们的方法聚焦于局部频谱特性，从而在增强对细微异常的敏感性的同时，减轻全局变换中固有的细节损失。

我们识别出流行的主干网络架构在频域分析中的局限性。具体来说，标准的 2D-CNN 主干不足以捕捉图像块内 DCT 系数分布之间的深层依赖关系，并且缺乏有效进行块间频率分析的表征能力。此外，许多现有方法依赖于插值操作（例如，sinc 或双线性插值）来对齐输入维度，这通常会削弱频域伪影，并降低判别性能。

受到这些观察结果的启发，我们在频率分析流程中提出一个本地-全局频率框架。该框架集成了两个互补分支：本地分支：我们设计了一种块内/块间多尺度频率卷积网络，以层次化地建模块内光谱系数相关性和块间过渡模式，从而实现局部频率特征提取。全局分支：我们引入尺度不变微分分析 (SIDA)，应用于原始分辨率的感兴趣区域裁剪。这种方法避免了由于重采样导致的高频降解，并保留了对全局模式分析至关重要的尺度不变光谱指纹。

此外，我们还通过利用一个在图像分类任务上预训练并在深伪检测任务上微调的 EfficientNet 骨干网络，结合空间域分析管道，以捕捉局部空间伪迹，如纹理、边缘和结构不一致性。

为了在频域和空间域之间架起桥梁，我们进一步提出了一个分层的跨模态融合框架。在浅层中，我们引入了一个频率感知注意力增强模块 (FAAE)，它从光谱特征生成逐块的、频率定位的注意力图。这通过突出局部伪造线索来增强空间表示。在深层中，我们实现了一个混合跨模态注意力融合模块，它结

合了用于空间-频率一致性的混合注意机制和用于异常放大的差分增强门控。这些模块共同通过分层交互实现多模态表示的有效整合。

3.2 局部-全局频率框架

形式上，令输入图像表示为 $X \in \mathbb{R}^{C \times H \times W}$ ，其中 H 和 W 分别表示高度和宽度， C 表示通道数。根据 [?] 的方法，我们首先将 RGB 图像转换为 YCbCr 颜色空间。随后，对所有三个通道应用块大小为 8×8 的块状 DCT 变换，这与标准图像压缩协议如 JPEG 一致。然后，使用之字形扫描将每块 DCT 系数展开成一维向量，该扫描按系数的低到高频顺序排列。这一过程将跨通道的相同频率分量的系数对齐，并系统地基于它们的频率内容组织所有块。结果表示为一个重构的四维张量 $\tilde{X} \in \mathbb{R}^{C \times 64 \times H/8 \times W/8}$ ，该张量明确地保留了对后续分析至关重要的空间-频率关系。如图 2 所示。

3.2.1 全局分支 传统的预处理流程通常将输入图像调整为固定尺寸，以方便批处理。然而，这种操作不可避免地改变了伪造区域的光谱能量分布，并破坏了高频信息的完整性。为了解决这一局限性，整体分支引入了 SIDA 模块，该模块在保持光谱特征形状一致性的同时，保留了原始分辨率下的分块 DCT 变换，适用于具有不同空间维度的图像。具体来说，我们首先使用人脸跟踪算法定位面部区域，以获得边界框坐标，并在这些区域内专门执行分块 DCT 变换，以保留区域特定的光谱信息。

对于块间尺度分析，本地化操作通常会导致相邻行和列块之间频率分布的突然变化。为了捕捉过渡性不一致，我们应用水平和垂直差分运算来消除语义信息，同时放大特定于操作的异常。我们在频率变换特征 $\tilde{X}$ 上计算块间差分运算，从而得出块间行差分特征 $\tilde{X}_{row} \in \mathbb{R}^{C \times 64 \times (H/8) \times ((W/8)-1)}$ 和块间列差分特征 $\tilde{X}_{col} \in \mathbb{R}^{C \times 64 \times ((H/8)-1) \times (W/8)}$ ，其可以用公式 1 和 2 表示。

$$\tilde{X}_{row} = \tilde{X}_{:, :, m+1, :} - \tilde{X}_{:, :, m, :}, m \in [1, (H/8) - 1] \quad (1)$$

$$\tilde{X}_{col} = \tilde{X}_{:, :, :, n+1} - \tilde{X}_{:, :, :, n}, n \in [1, (W/8) - 1] \quad (2)$$

在块内尺度上，伪造区域通常显示异常的光谱能量分布。块内系数差异量化了相邻频率分量之间的局部相关性。然后对 $\tilde{X}$ 进行块内差分操作以获得块内差分特征 $\tilde{X}_{intra} \in \mathbb{R}^{C \times 63 \times (H/8) \times (W/8)}$ ，如方程 3 所示，并在块内维度上进行零填充以保持维度一致性。

$$\tilde{X}_{intra} = \tilde{X}_{:, l+1, :, :} - \tilde{X}_{:, l, :, :}, l \in [1, 63] \quad (3)$$

随后，对三个导数特征图的绝对值计算四个统计指标——均值、标准差、偏度和峰度，如方程 4-7 所示：

$$Mean_k = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |\tilde{X}_k^{i, :, i, j}|, (k = row, col, intra) \quad (4)$$

$$Std_k = \sqrt{\frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (|\tilde{X}_k^{i, :, i, j}| - Mean_k)^2} \quad (5)$$

$$Skew_k = \frac{\frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (|\tilde{X}_k^{i, :, i, j}| - Mean_k)^3}{Std_k^3} \quad (6)$$

$$Kurt_k = \frac{\frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (|\tilde{X}_k^{i, :, i, j}| - Mean_k)^4}{Std_k^4} \quad (7)$$

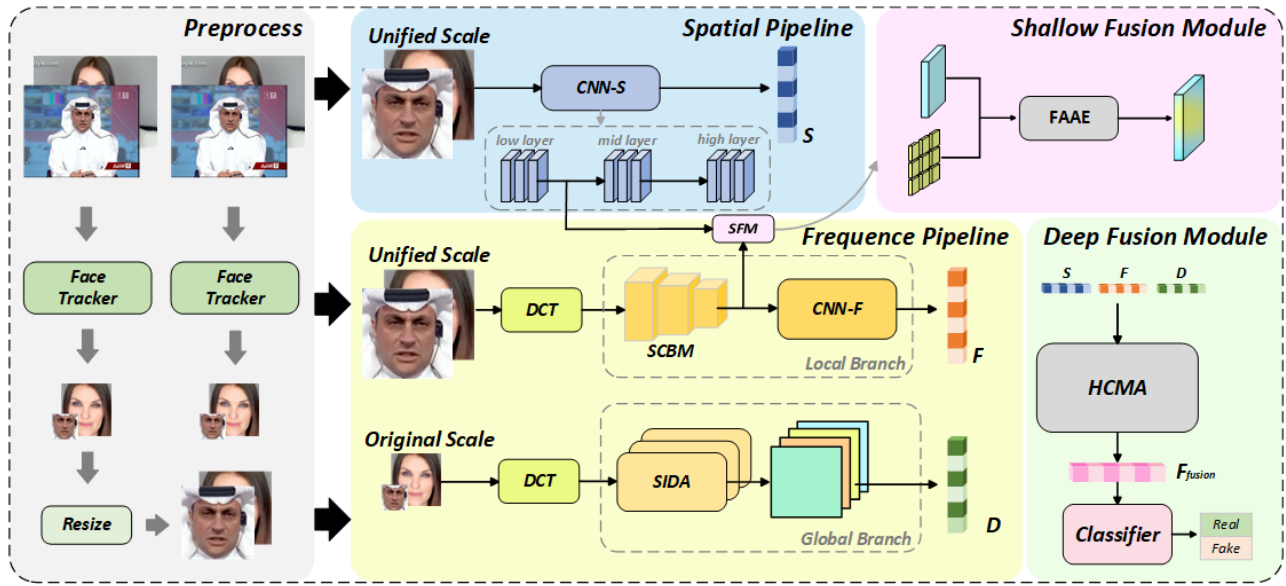


Figure 1: SFCL-HCMF 的架构。

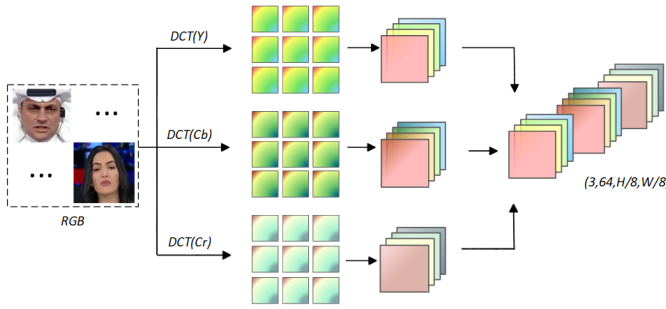


Figure 2: 分块 DCT 变换。

这个聚合过程将局部的差异强度转化为全局统计描述符，从而捕捉伪造痕迹的整体分布模式，而不是依赖于特定位置的伪影。随后，四种类型的差异统计特征被展平并连接以形成如方程 8 所示的全局差异频率表示 $D \in \mathbb{R}^{2304}$ ，该表示有效地编码了垂直、水平和光谱内异常模式。

$$D = \text{cat}(\text{flat}(\text{Mean}_k), \text{flat}(\text{Std}_k), \text{flat}(\text{Skew}_k), \text{flat}(\text{Kurt}_k)) \quad (8)$$

3.2.2 局部分支。2D CNNs 在处理高维光谱特征时表现出固有的局限性，未能捕捉单个图像块内的交叉频率相关性（沿着通道维度）。为了克服这些局限性，我们设计了块内/块间多尺度频率卷积网络，它由 SBCM 和 CNN-F 组成。局部分支的流程如图 3 所示。

二维卷积神经网络在处理高维光谱特征时表现出固有的局限性，因为它们无法捕捉单个图像块中跨频率关联（即沿通道维度）。为克服这些挑战，我们提出了一种多尺度频率卷积网

络，该网络由光谱带卷积模块（SBCM）和频率卷积神经网络（CNN-F）组成。

SBCM。本模块采用堆叠的 3D 卷积块，核大小 $= (x, 1, 1)$ ，以捕捉光谱系数之间的通道间相关性。通过在光谱带维度上滑动 3D 卷积核，SBCM 逐步学习邻近频带之间的局部交互模式。初始层使用较大的卷积核来建立低、中、高频分量之间的长距离依赖关系，以实现跨频带特征聚合；中间层逐步减小 x ，以专注于由伪造伪影引起的突发光谱转换，从而增强中频区域的异常响应；最终层采用最小的 x ，以隔离高频噪声模式并在关键光谱带中放大分布的不规则性。

处理后的光谱-语义特征，记为 $X_f^{local} \in \mathbb{R}^{C \times 3 \times (H/8) \times (W/8)}$ ，其中 $C=64$ ，从而封装了层次上精炼的频率特定表示。

我们随后将张量的前两个维度进行展平，这与初始的直接展平方法不同，此时每个空间位置的 192 维特征现在包含了具有跨波段相关性的高级语义特征，如公式 9 所示。

$$X_f^{local} = \text{flat}(X_f^{local}, \text{dim} = (0, 1)) \quad (9)$$

CNN-F 为了对异构图像块之间的关系进行建模，我们通过修改输入层来调整传统的 CNN 架构——以 Xception 为例。具体而言，我们去掉了前两个卷积层和一个深度可分离卷积块，用 SBCM 替换它们。为了保持架构的兼容性，然后我们调整后续深度可分离卷积块的输入通道数，以匹配输出光谱特征 X_f^{local} 的通道维度。这个层次化的处理框架使得能够建模 DCT 系数之间的块内和块间相关性，从而产生频率感知特征 $F \in \mathbb{R}^{2048}$ 。

我们旨在放大空间域模型浅层中的局部伪造伪迹，从而增强微小操作痕迹的显著性，以促进判别特征表示的学习。频率感知注意力增强模块的结构如图所示 ??。

该模块最初接收两个输入：浅层空间特征 X_S 和 SBCM 输出的光谱特征 X_f^{local} 。我们首先通过并行卷积投影生成跨模态查询-键对，同时对齐通道维度以确保兼容性，如公式 10 和 11

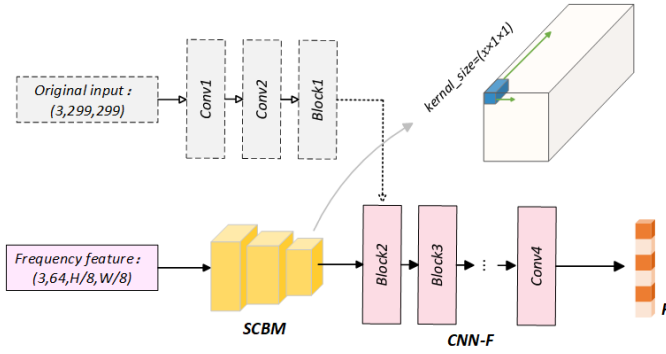


Figure 3: 局部分支的流程（块间/块内多尺度频率卷积网络）

所示。

$$M_{query} = cat(Conv_q^F(X_f^{local}), Conv_q^S(X_S)) \quad (10)$$

$$M_{key} = cat(Conv_k^F(X_f^{local})^T, Conv_k^S(X_S)^T) \quad (11)$$

随后通过串联键对之间的缩放点积计算得到跨模态注意力权重矩阵 α ，该矩阵通过关联频率局部化模式与空间异常来封装光谱-空间交互动态，如公式 12 所示。

$$\alpha = Softmax(\frac{M_{query}M_{key}^T}{\sqrt{2H}}) \quad (12)$$

空间细化分支通过注意力门控注入频率上下文约束。这里， $Conv_F$ 将光谱域特征投影以与空间特征维度对齐， γ_S 表示可学习参数，动态调节光谱特征对空间分支的贡献强度， σ 代表 Sigmoid 激活函数， $Conv_S$ 调整频率感知注意力矩阵的维度，以确保跨模态特征融合期间的维度兼容性，如公式 13 所示。

$$Y_S = X_S + BN(Conv_S(\alpha \odot Conv_F(X_f^{local}) \cdot \sigma(\gamma_S))) \quad (13)$$

为了在空间和光谱域之间建立深入的交互推理，我们提出了一个 HCMA 多模态融合模块，如图 4 所示。该模块首先通过线性变换层 $S', F' \in \mathbb{R}^{1792}$ ，将空间深度特征 $S \in \mathbb{R}^{1792}$ 和块内/块间频率相关特征 $F \in \mathbb{R}^{2048}$ 投射到一个统一的嵌入空间，从而消除模态特定的维度差异。

使用空间特征作为查询向量 $Q = S'W_Q$ ，频率特征作为键值对 $K = F'W_K, V = F'W_V$ ，其中 $W_Q, W_K, W_V \in \mathbb{R}^{1024 \times 1024}$ 是可学习的参数矩阵。然后，使用多头注意力捕捉跨模态的全局依赖关系，以生成注意加权表示，如公式 14 中所示。

$$A = softmax(\frac{QK^T}{\sqrt{1024/h}})V \quad (14)$$

为了增强融合性能，我们通过将注意力输出与经过一维卷积处理的空间特征相加并进行批归一化来引入残差连接，从而保持梯度稳定性，如方程 15 所示。

$$A_{res} = A + BN(Conv_{1 \times 1}(S')) \quad (15)$$

考虑到差分统计特征对被操纵区域的敏感性，设计了一种动态门控机制——通过 Sigmoid 激活将差分频率统计表示 $D \in \mathbb{R}^{2304}$ 投影到门控系数上，该机制调整特征以增强异常响应，如方程 16 和 17 所示。

$$g = \alpha(W_g D + b_g), W_g \in \mathbb{R}^{1024 \times 2304} \quad (16)$$

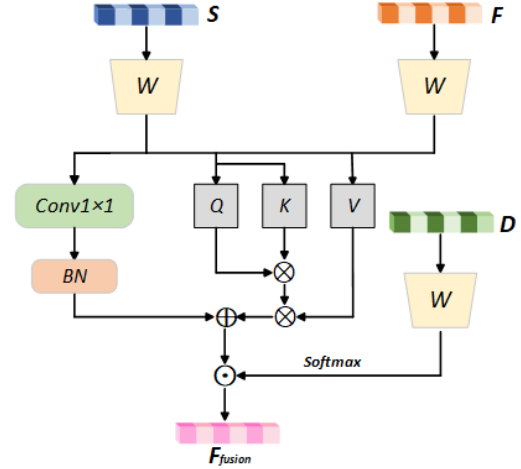


Figure 4: 混合跨模态注意力融合的结构。

$$F_{fused} = A_{res} \odot g \quad (17)$$

该架构实现了分层的跨模态特征协同，结合了全局依赖建模与局部异常增强能力。最终，融合后的多模态语义表示向量被输入到分类器中，以执行用于识别深度伪造检测任务的判别性决策。

4 实验

对于每个视频片段，我们首先提取相同数量的帧。然后，我们使用最先进的人脸检测器 dlib 来记录每张图像的人脸边界框坐标。在预处理过程中，基于这些坐标（面部追踪器）裁剪面部区域，并应用块大小为 8×8 的块状 DCT 变换。空间管道和频域管道的局部流接受解析度为 380×380 的输入，而频域管道的全局流通过 SIDA 操作将数据处理为相同的尺寸。在模型架构中，我们采用在 ImageNet 上预训练的 EfficientNet-B4 作为空间域管道的骨干。在 SCBM 中，3D 卷积核的大小依次设为 7、5 和 3。HCMA 模块利用 8 头的多头注意力。为了训练，使用 Adam 优化器进行模型优化，学习率为 0.001，权重衰减为 $1e-8$ ，批量大小为 20。模型在两块 NVIDIA RTX 4090 GPU 上训练 20 个周期。

Table 1: 在 FF++ 数据集中，针对不同压缩级别的内部数据集评估结果。最佳结果用粗体标出，次佳结果用下划线标出。

Methods	FF++(HQ)		FF++(LQ)	
	Acc(%)	AUC(%)	Acc(%)	AUC(%)
Xception[?]	95.73	96.30	86.86	89.30
X-Ray[?]	-	87.40	-	61.60
Two Branch[?]	-	98.70	-	86.56
SPSL[?]	91.50	95.32	81.57	82.82
MAD*[?]	97.17	99.28	87.35	89.92
EDD[?]	96.19	98.81	88.69	<u>91.27</u>
M2TR*[?]	96.73	99.16	<u>89.36</u>	91.08
HFI-Net[?]	95.12	98.66	86.90	90.75
ED[?]	94.98	98.30	-	-
SPCL-HCMF(ours)	97.43	99.58	91.77	94.21

4.1 数据集

我们的实验主要在 FaceForensics++ (FF++) [?] 数据集上进行。FF++ 包含 1000 个原始视频和 4000 个由四种不同的伪造方法生成的对应操控视频：Deepfakes、Face2Face、FaceSwap 和 NeuralTextures。该数据集包括三个压缩级别：c0（原始/未压缩）、c23（轻压缩）和 c40（重压缩）。我们专注于 c23 和 c40 版本进行评估。

Celeb-DF [?] 数据集包含两个版本（v1 和 v2），其中有 59 位名人进行的访谈，覆盖不同年龄、种族和性别的人群。v1 子集包括 409 个真实视频和 795 个合成视频，而 v2 将这一数量扩展至 590 个真实视频和 5,639 个被操控的视频。我们使用增强版本的 v2 进行了跨数据集评估，它结合了更多样化的操控类型。

DFDC（DeepFake Detection Challenge）[?] 数据集是一个基准数据集，包括来自 3426 名演员的超过 10 万个视频片段。在我们的实验中，DFDC 数据集主要用于评估我们模型的跨数据集泛化能力。

我们采用准确率 (Acc) 和 ROC 曲线下面积 (AUC) 作为评价指标，这是深度伪造检测研究中的标准度量。所有实验结果均在图像层面报告。

为了评估所提方法的有效性，我们首先在 FaceForensics++ (FF++) 数据集上进行数据集内实验。按照标准协议，该数据集由 1000 个视频组成，其中 720 个用于训练，140 个用于验证，140 个用于测试。对于每个视频，提取 270 帧，并应用四倍数据增强以平衡类别分布。

我们在高质量 (HQ, c23) 和低质量 (LQ, c40) 压缩设置下进行实验，并在表 1 中报告准确率 (Acc) 和曲线下面积 (AUC) 指标。我们的方法与最新的检测方法进行基准比较，包括 X-Ray [?]、MAD [?] 和 M2TR [?] 等。如结果所示，我们的方法在两种压缩水平上都取得了优越的性能。特别是，与诸如 SPSL [?] 和 HFI-Net [?] 等频率感知基线相比，我们的方法在 HQ 上的检测准确率分别提升了 5.9 % 和 2.3 %。

4.2 跨数据集评估

为了证明我们方法的泛化能力，我们在多个基准数据集上进行了跨数据集评估。由于数据集之间存在显著分布转移，此设置提出了更大的挑战。具体来说，我们在 FF++ (HQ) 上训练我们的模型，并在 Celeb-DF (v2) 和 DFDC 上进行评估。对于每个视频，我们提取 50 帧并计算帧级 AUC 分数。如表 2 所示，我们的方法在所有跨数据集评估中都取得了优异或可比的性能。值得注意的是，像 F^3 -Net [?] 这样的频率感知方法在跨数据集的适应性方面存在限制。相比之下，我们的方法在 Celeb-DF (v2) 上分别超过它们 9.5 % AUC。与空间频率融合方法 Two-branch [?] 和 SFDG [?] 相比，我们的方法表现出更好的鲁棒性，证明了所提出的层次跨模态融合策略在利用空间和频率特征的互补优势方面的有效性。

在本节中，我们对提出的关键组件——SCBM、SIDA 和 HCMF 进行消融研究，以系统地研究它们对检测性能的贡献。具体来说，我们依次移除每个组件，并在 FF++ (c23) 上评估相应的 AUC 分数，结果详见表 ??。

对 SCBM 的消融研究。SCBM 旨在捕捉个别图像块中细微的基于频率的伪造痕迹。为了评估其有效性，我们从频域的局部分支中移除了 SCBM，同时保持其他所有模块不变，直接将频域特征输入到 CNN-F 中。正如表 ?? 所示，这一修改导致性能下降了 0.5 %，证实了 SCBM 捕捉 DCT 系数突变并增强局部光谱特征表示的能力。

关于层次化跨模态融合的消融研究。所提出的层次化融合框架由 FAAE 模块和 HCMA 模块组成，这两个模块共同对空间和多尺度频域特征进行建模和整合。将这种机制替换为简单的空间和频率特征连接，导致性能下降了 0.6 %，证明了我们设计的有效性。进一步的消融研究通过分别去除 FAAE 和 HCMA 模块进行。如表 ?? 所示，只有在联合使用这两个模块时，框架才能实现最佳性能，突出了它们在增强跨模态特征融合中的互补作用。

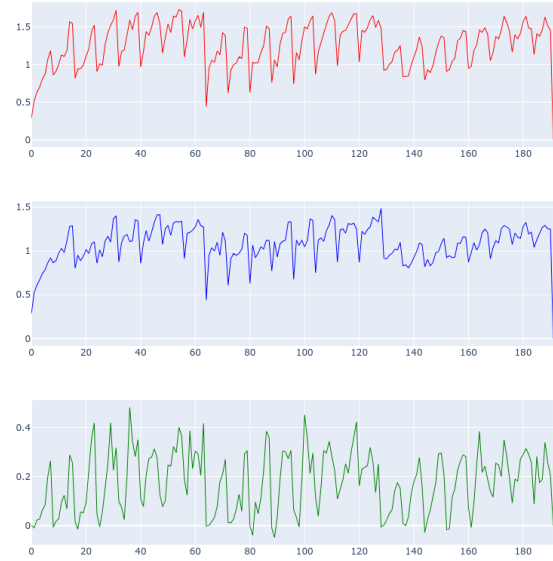


Figure 5: 在 SIDA 中差分统计指标的折线图。红色代表真实人脸，蓝色表示伪造人脸，绿色表示它们之间的差异。

对 SIDA 的削弱实验。SIDA 旨在充分利用全局频率微分统计特征。通过去除频域中的 SIDA（全局局部分支），仅保留空间特征和局部频域特征的层次跨模态融合。如表 ?? 所示，性能下降了 1.4 %。这验证了 SIDA 在特征精炼方面的有效性。

为了提供对 SIDA 效果更加直观的可视化，我们绘制了折线图（使用平均值作为示例，并且仅考虑 Y 分量），用于展示 SIDA 中的差异统计指标。如图 5 所示，192 维特征空间分为三个部分：前 64 维对应于块间行差异，中间 64 维代表块间列差异，最后 64 维表示块内差异。结果表明，SIDA 在中高频范围内有效地增强了真实和伪造图像之间的差异。

在本文中，我们引入了一种具有分层跨模态融合的空间-频率协同学习框架，用于鲁棒的深度伪造检测。首先，提出了一种基于块的 DCT 分析与块内/块间多尺度频率卷积网络，以增强局部表示学习并在频率域中实现精确的伪造定位。其次，设计了一种尺度不变的差分分析方法，以捕获整体伪造痕迹，同时减轻多尺度差异和后处理干扰。最后，分层跨模态融合机制统一了浅层注意力优化和深层动态调制，解决了跨模式的空间-频率相互依赖性和特征异质性。广泛的同数据集和跨数据集实验展示了我们方法的优越检测性能和泛化能力。消融研究进一步验证了我们设计中各种关键组件的关键作用。

Table 2: 不同方法在各种数据集上的性能比较。最佳结果以粗体显示，次佳结果以下划线标出。

Methods	Years	Training dataset	FF++(c23) AUC(%)	Celeb-DF(v2) AUC(%)	DFDC AUC(%)
Xception[?]	2017	FF++(c23)	96.30	65.30	72.20
Two-branch[?]	2020		93.18	73.41	-
X-Ray[?]	2020		87.40	-	70.00
F ³ -Net[?]	2020		98.10	65.17	-
MAD*[?]	2021		99.28	67.44	65.56
M2TR*[?]	2022		99.16	65.17	-
SFDG[?]	2023		99.53	75.83	<u>73.64</u>
DAW-FDD[?]	2024		98.28	74.42	61.47
SPCL-HCMF(ours)	-	FF++(c23)	99.58	<u>74.68</u>	73.71

