E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0120-18

论文引用格式:Long M, Yin X, Zhang L B and Peng F. 2026. Multi-branch network based on multi-domain feature fusion for Deepfake detection. Journal of Image and Graphics, 31(1):0120-0137(龙敏, 尹茜, 张乐冰, 彭飞. 2026. 基于多域特征融合的多分支网络用于Deepfake检测. 中国图象图形学报, 31(1):0120-0137)[DOI:10.11834/jig.240681]

基于多域特征融合的多分支网络用于Deepfake检测

龙敏[1],尹茜[2],张乐冰[3],彭飞[4*]

1. 广州大学电子与通信工程学院,广州　510006; 2. 长沙理工大学计算机与通信工程学院,长沙　410114;
3. 怀化学院计算机与人工智能学院,怀化　418000; 4. 广州大学人工智能学院,广州　510006

摘　要: **目的**　由于现有的基于卷积神经网络的检测方法往往局限于观察全局或局部时空特征,难以获取更全面的伪造线索,从而限制了检测方法的泛化能力。为了解决这一问题,本文提出一种基于多域特征融合的多分支网络框架(multi-branch multi-division,MBMD),综合利用频率域、空间域和时空域信息,以挖掘更全面细致的伪造线索。**方法**　在频率流中对图像进行DCT(discrete cosine transform)变换,去除低频分量并保留高频分量,以捕捉图像细微结构变化的频率特征。在空间流中,设计了空间特征增强块(spatial feature enhancement block,SEB)对CNN(convolutional neural network)的浅层特征进行多尺度增强,以捕捉图像中的局部异常区域。此外,在时空流中设计了信息补充块(information supplement block,ISB),将空间流中的局部特征与视觉Transformer捕获的全局高层特征相结合,使网络能够更全面地捕捉全局和局部的时空不一致。最后,通过交互融合模块(interactive fusion module,IFM)将频率域、空间域和时空域信息进行增强融合,以提取更全面细致的特征。**结果**　实验在不同数据集上与最新的方法进行了比较:在跨数据集实验中,相比于性能第2的检测模型,在Celeb-DF-v2数据集中ACC值提高2.63%,AUC值提高3.01%;在DFDC数据集中,相比于最新的检测模型,AUC值提高4.43%。同时通过消融实验分析了不同模块对泛化性能的影响,验证了提出方法的有效性。**结论**　在不同数据集上的实验表明,提出的方法在未知数据集上具有更好的泛化能力。

关键词: Deepfake检测;数字图像取证;多域特征融合;多分支;局部全局作用

Multi-branch network based on multi-domain feature fusion for Deepfake detection

Long Min[1], Yin Xi[2], Zhang Lebing[3], Peng Fei[4*]

1. School of Electronics and Communications Engineering, Guangzhou University, Guangzhou 510006, China;
2. School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China;
3. School of Computer and Artificial Intelligence, Huaihua University, Huaihua 418000, China;
4. School of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China

Abstract: **Objective** In recent years, the rapid development of deepfake technology has led to the proliferation of realistic manipulated videos on social media. Even non-expert users can easily manipulate facial content using various applications

收稿日期:2024-11-18;**修回日期**:2025-04-08;**预印本日期**:2025-04-15

*通信作者:彭飞　eepengf@gmail.com

基金项目:广州市科技计划项目(2025A03J3122);广东省自然科学基金项目(2025A1515010264)

Supported by: Science and Technology Projects in Guangzhou (2025A03J3122); Natural Science Foundation of Guangdong Province, China (2025A1515010264)

and open-source tools. Given the sensitivity of personal identity information, false content tends to spread rapidly on the internet, which raises public concerns regarding video authenticity and personal privacy security. Although existing deepfake detection methods perform well in specific scenarios, their performance often degrades considerably when confronted with the ever-changing conditions of real-world scenarios. In other words, detection performance degrades when the forgery technique is unknown, which severely affects the performance of most deepfake detection methods in real-world applications. Furthermore, given that the existing deepfake detection methods based on convolutional neural networks often focus on either global or local spatiotemporal features, they still exhibit complexity in capturing global features and temporal information. The difficulty in capturing comprehensive forgery clues limits the generalization capability of detection methods. To address these issues, a multi-branch network based on multi-domain feature fusion is proposed in this study. This network comprehensively utilizes frequency, spatial, and spatiotemporal domain information to mine more detailed and comprehensive forgery clues. **Method** In the frequency stream, the discrete cosine transform is applied to the images, the low-frequency components that introduce additional noise and information redundancy are removed, and the high-frequency components are retained to capture frequency features of subtle structural changes in the images. In the spatial stream, a spatial feature enhancement block is designed to enhance the shallow features of CNN at multiple scales, and local abnormal regions in the images are captured. In the spatiotemporal stream, the Vision Transformer is used to process the frame sequence, and the global high-level features are obtained to perceive global spatiotemporal information. An information supplement block is developed to combine local features from the spatial stream with global high-level features captured by the Vision Transformer. As a result, the network can comprehensively capture global and local spatiotemporal inconsistencies. Finally, an interactive fusion module is used to enhance and fuse information from the frequency, spatial, and spatiotemporal domains to extract more detailed and comprehensive features for deepfake detection. **Result** A systematic comparison was conducted between the proposed method and the state-of-the-art approaches across multiple datasets. In cross-dataset generalization experiments, the advantages of the proposed method were even more pronounced: on the Celeb-DF-v2 dataset, compared to the second-best performing detection model, the proposed method improved ACC by 2.63% and AUC by 3.01%; on the more challenging DFDC dataset, compared to the latest detection model, the proposed method achieved a notable improvement of 4.43% in AUC, highlighting its superior cross-domain adaptability. To further investigate the contribution of the model design to generalization performance, a systematic ablation study was conducted, which thoroughly analyzed the impact of different modules in cross-dataset scenarios. The experimental results clearly validate the effectiveness of each key component and explain the mechanism through which the proposed method enhances generalization capability from a quantitative perspective. This comprehensively confirms the innovativeness and practical utility of the proposed approach. **Conclusion** The proposed method has been validated through systematic comparison with state-of-the-art approaches on multiple datasets. It not only achieves advanced performance levels but also demonstrates excellent cross-domain adaptability and interpretability, providing an effective solution for research in related fields.

Key words: Deepfake detection; digital image forensics; multi-domain feature fusion; multi-branch; local-global role

0 引　言

随着变分自动编码器和生成对抗网络(generative adversarial network，GAN)等深度生成模型的快速发展，视觉伪造现象在社交媒体和互联网上变得异常普遍。即使对非专业用户而言，利用各种应用程序和开源工具也能轻松进行面部内容的操纵。最新的伪造技术通常称为深度伪造，即Deepfake，这种技术利用生成网络生成高质量的图像或视频，产生逼真的面部效果(丁峰 等，2024)。尽管Deepfake技术最初用于正当目的，但也存在着被不法分子滥用的风险。由于个人身份信息的敏感性，虚假内容往往会在网络上迅速传播，引发严重的信任和安全问题。因此，迫切需要相应的防御方法来检测这种具有潜在危害的伪造视频。

考虑到这项技术对社会造成的严重影响，研究人员已经提出多种方法来检测视频中的虚假面部。尽管现有的检测方法在处理深度伪造任务时取得了显著进展，但深度伪造技术的快速发展带来了严峻

的挑战：现有网络在未经重新训练的情况下对新的伪造样本的检测性能通常较差，即当伪造技术未知时，检测性能会下降。这种性能下降主要是因为未能提取出伪造品内在痕迹的特征。由于大部分方法是在相对同质的数据集上进行训练和测试的，基于卷积神经网络（convolutional neural network，CNN）的检测器可能存在过拟合问题，因为它们过度依赖于训练数据的特征分布，而未能捕捉真实的伪造痕迹。在这种情况下，经过训练的检测器在数据集内的检测效果非常优异，但在交叉数据集的检测中性能通常会显著下降，因为它们无法有效地适应新的数据分布，这使得大多数深度伪造检测方法在实际应用中的性能受到了严重影响。因此，提升模型的泛化能力成为现有深度伪造检测系统的主要关注点之一。

为了提高泛化能力，一些基于图像的检测方法重点关注在人脸伪造中特别突出的异常结构，例如纹理特征。早期的伪造操作通常在低频信息上留下更多的伪造痕迹，因此很容易被人眼识别。然而，随着人脸伪造技术的进步，最新的伪造方法生成的数据难以辨别。为此，研究人员引入了频域模态，并综合所有频域信息进行伪造检测（Qian等，2020；Frank等，2020）。然而，直接使用所有频域信息可能带来大量冗余信息，从而阻碍模型快速感知伪造操作留下的痕迹。另一方面，一些研究人员考虑将频域信息与空间信息相结合。对单个图像的频域信息进行编码（Chen等，2021；Gu等，2022a），但是未充分利用频域和RGB域的信息，难以进行高效的特征嵌入，并且粗糙的模态变换也会导致信息的丢失。这些方法实现了频域信息与空间信息的结合，但是在频率信息选择性使用上仍存在冗余，影响了模型的效率。因此，如何有效利用频率信息，减少冗余并突出伪造痕迹，仍是亟待解决的问题。随着视频伪造技术的发展，基于视频的检测方法（Güera和Delp，2018；Sabir等，2019；Masi等，2020；Chintha等，2020；Ganiyusufoglu等，2020；Zhang等，2021）发现，虚假人脸视频通常是通过逐帧操作生成的，无法避免时间不一致，例如面部异常抖动。因此，它们直接利用了循环神经网络（recurrent neural network，RNN）、长短期记忆网络（long short-term memory，LSTM）或三维卷积神经网络（3D convolutional neural network，3DCNN）来捕获时间线索。Hu等人（2022a）将检测转化为未来帧推断任务，以探索视频数据中的时间特征并感知异常的时间特征，但它无法感知不同时间跨度的时间异常。Zhu等人（2024）利用频率信息来检测相邻帧中的光照不一致性。然而，这些方法在挖掘重要的动态不一致性方面受到限制。目前，主流的基于视频的检测方法利用卷积神经网络（CNN）从全局或局部区域来捕获存在的时空不一致性。Gu等人（2022b）通过由连续帧组成的“视频片段”深入研究时间局部性来捕获时空信息。Haliassos等人（2021）则利用针对性的高层次语义不规则的嘴部运动，使用唇读网络进行伪造检测。Yu等人（2023）利用CNN网络从全局角度出发来关注不同时间段的局部时空异常。由于CNN在处理单幅图像时存在优势，擅长处理局部特征和空间信息，使得这些检测方法在泛化方面有所提升，但其对全局特征和时间信息的捕获仍然不足。最近的研究将Transformer扩展到计算机视觉（computer vision，CV）任务，由于其对建模远程和全局上下文信息的强大能力，一些基于Transformer的检测方法也相继提出（Zheng等，2021；Yang等，2023）。尽管上述方法取得了一些进展，但是利用Transformer捕获的特征未能巧妙结合局部和全局信息之间的关联，使得检测方法的泛化能力仍然有限。

随着技术的不断进步，研究者逐渐认识到单一信息域的依赖存在局限性，特别是在Deepfake视频检测任务中，频域、空域与时域信息的有效融合已成为提升检测准确性和鲁棒性的关键方向。虽然近年来许多研究已尝试结合多模态信息进行Deepfake检测，但仍存在冗余信息处理不当和频率信息未被高效利用的问题。例如，许多利用频域的检测方法未能从频率信息中选择性地提取有用的高频分量，而是处理了包含大量噪声的全频域信息，从而影响了检测效果。此外，现有的方法在结合局部与全局特征方面也有一定局限性。大多数方法依赖于单一网络进行特征提取，无法有效地发挥不同网络在捕捉局部和全局时空信息方面的优势。根据现有研究，卷积神经网络（CNN）在捕捉局部空间信息方面表现出色，而Transformer网络则在捕捉全局时空信息时具有独特优势。Jia等人（2022）指出，真假人脸在频域中的分布存在显著差异，特别是虚假人脸在高频段的能量要比真实人脸更加丰富。通过上述分析，本文提出一种基于多域特征融合的多分支网络，用

于深度伪造检测。与现有方法不同，本研究通过整合频域、空间域和时空域的信息，结合视觉 Transformer 和 CNN 的优势，全面挖掘和利用不同信息域中的检测线索。

本文主要贡献归纳如下：1）提出的框架整合频率域、空间域和时空域信息，充分发挥频率信息在细节捕捉中的优势，以及空间与时空信息在捕捉局部异常和全局一致性中的作用，并通过设计的交互融合模块（interactive fusion module，IFM），实现各信息域之间的融合。通过多域信息的整合，模型能够更全面、细致地捕捉伪造痕迹，提高 Deepfake 检测的泛化能力。2）提出的空间特征增强块（spatial feature enhancement block，SEB），通过多尺度增强融合实现了对空间流中的局部异常的精准捕捉，提升了对局部细节特征的敏感性。与此同时，设计的信息补充块（information supplement block，ISB）能够帮助时空流同时捕捉全局和局部时空异常，弥补视觉 Transformer 在捕捉局部时空异常方面的不足。3）在 FF++、Celeb-DF 和 DFDC 数据集上进行广泛实验，证明了提出的方法在人脸伪造视频检测方面比现有方法具有更佳的泛化能力。

1 相关工作

1.1 Deepfake 生成技术

Deepfake 是一种基于深度学习的人脸操纵技术，能够对图像和视频中的人脸进行有目的地操纵，从而创建极其逼真的假象。Deepfake 生成大致可分为 4 类：重现、替换、编辑和合成。人脸重现包括保留源主体的身份，同时对主体的嘴巴、表情和身体进行处理，例如 Recycle-GAN（Bansal 等，2018）和 NeuralTextures（Thies 等，2019）。人脸替换涉及将源面孔的身份转移到目标面孔上。Faceswap 是一种典型的替换工具。Li 等人（2019）提出一种两阶段网络，利用启发式误差验证细化网络实现高质量的人脸替换。Kim 等人（2017）利用循环一致性 GAN 来交换源和目标的身份，同时保持目标的面部表情。人脸编辑和合成用于添加或删除人的属性，如修改年龄、性别或种族，如 Transeditor（Xu 等，2022）和 STIT（stitch it in time）（Tzaban 等，2022）。目前，各种开源工具和商业应用都可用于 Deepfake 生成。然而，这项技术也存在被恶意滥用的风险，这推动了 Deepfake 检测方法的开发和应用。

1.2 Deepfake 检测方法

1.2.1 基于 CNN 的 Deepfake 检测方法

基于 CNN 的 Deepfake 检测方法包括基于图像的检测方法和基于视频的检测方法。

1）基于图像的检测方法。在早期阶段，通常使用手工制作的特征和相机特征来捕捉伪造线索。然而，随着伪造技术的发展，这些常规特征已不足以有效应对越来越逼真的假人脸。随着深度学习的快速发展，基于 CNN 的检测方法应运而生，开启了对空间域伪造信息的探索。Chen 等人（2023a）将 Deepfake 检测表述为一个多任务学习问题，通过双粒度伪造进行全局人脸取证。部分研究利用 CNN 提取局部特征的优势，试图检测局部空间区域的非自然人脸特征，以发现各种伪造方法留下的细微瑕疵。Zhao 等人（2021a）通过使用多注意图捕捉具有辨别力的局部区域，将 Deepfake 检测转化为细粒度分类。Zhao 等人（2021b）利用真假图像具有不同源特征的特点，采用成对自一致性学习（pairwise self-consistent learning，PCL）来计算不同区域特征图的两两相似性。Li 等人（2023）利用超分辨率算法中部分特征恢复和增强的思想，设计了一种数据增强机制，帮助网络有针对性地恢复和增强局部空间特征，以提高低质量图像的检测性能。Shang 等人（2021）提出像素级关系模块和区域级关系模块来检测图像像素和局部区域之间的不一致性。然而，对于完全合成的人脸图像，该方法检测精度有所降低。此外，一些研究（Ganguly 等，2022；Coccomini 等，2022）建议利用局部和全局空间信息的互补性，以形成更全面的考虑。然而，这些方法都忽略了其他域的伪造信息，因此泛化能力仍然有限。

由于深度生成模型的上采样步骤会引入频域异常，如频率成分的分布差异和棋盘伪影，这启发了研究人员对频域伪造痕迹的探索。Durall 等人（2019）使用离散傅里叶变换（discrete Fourier transform，DFT）对幅度频谱进行平均，并利用分布差异来区分伪造图像。Frank 等人（2020）通过离散余弦变换（discrete cosine transform，DCT）将图像从空间域转换到频率域，并分析了各种 GAN 中的频率异常。虽然这些方法能有效检测完全由 GAN 合成的图像，但在处理小范围篡改和压缩图像时，其性能明显下降。Qian 等人（2020）通过使用频率感知分解和局部频率

统计来挖掘频域中的细微伪造模式。它解决了部分自适应频域特征学习的问题，但仍然依赖人类先验来划分频段。此外，局部篡改痕迹不会在全局频域造成明显异常，这促使人们考虑结合空间和频域来提取互补特征。Afchar等人（2018）基于图像的中观特性引入了两种CNN架构来检测假人脸。Wang等人（2022a）通过放大真假人脸在空间域和频率域的局部差异来检测伪造人脸。Luo等人（2021）利用高频噪声和低频纹理来揭示伪造痕迹。Liu等人（2021）认为上采样结构在相位频谱中造成的像素差异比在幅度频谱中造成的差异更显著，通过利用相位信息和浅层特征，并抑制高层语义信息使网络聚焦于纹理丰富的局部区域。Chen等人（2022）提出两阶段学习框架，用于学习更具代表性的特征，并引入用于人脸变换和重现的AFS（adjustable forgery synthesizer）。它使用小波分解来模拟合成Deepfake图像的过程，以防止过度拟合。虽然这些方法都取得了良好的性能，但其泛化能力仍有待提高。值得注意的是，基于图像的检测方法通常会忽略时间线索的重要性。

2）基于视频的检测方法。与静态面部图像相比，描述面部特征运动的动态信息更难以在不留下视觉痕迹的情况下伪造。由于Deepfake视频是逐帧生成的，这会出现帧间不一致性。因此，大多数研究都集中在Deepfake视频检测的时间不一致性上。早期的研究直接使用3D CNN提取时空信息，也有一些方法先提取帧级CNN特征，然后使用递归神经网络、光流（Amerini等，2019）或长短期记忆（LSTM）（Chintha等，2020）提取时间线索进行伪造检测。Xia等人（2022）提出像素级映射、块级映射和区域级映射，利用帧间颜色成分的差异提取真假视频的纹理特征。祝恺蔓等人（2022）提出利用关键帧提取人脸的空间特征和帧间特征，并融合多个关键帧的人脸特征信息，来实现针对人脸交换的伪造检测。然而，这些方法并非专门为Deepfake检测设计，提取的时间线索也比较粗糙和有限。目前，Deepfake视频检测方法（Ismail等，2022）侧重于全局时空不一致性。Li等人（2018）利用眨眼频率的异常变化进行检测。Gu等人（2021）引入用于Deepfake视频检测的STIL（spatial-temporal inconsistency learning）块，并利用沿两个正交方向的时间差捕获了细粒度的时间不一致性。Li等人（2020a）设计了一种新的时空实例来捕捉细微的不一致性。Hu等人（2022b）基于帧级和与时间相关的残差特征检测压缩Deepfake视频。Chen等人（2022）通过时空注意力机制和ConvLSTM有效利用了帧间和帧内信息，从而提高检测性能。Yu等人（2022）提出一种提高泛化能力的共性学习策略，它假定伪造方法会在视频中留下一些相似的伪造痕迹。Pang等人（2023）提出一种多速率激励网络，利用瞬时不一致激励模块和长期不一致激励模块来有效感知帧内和帧间动态时空不一致。然而，这些方法更侧重于全局区域，对局部区域的关注不足。因此，研究人员开始探索局部和全局的时空视图以捕获更全面的时空信息。Hu等人（2021）提出DIANet（dynamic inconsistency-aware network），用于捕捉全局和局部帧间的不一致性，以检测虚假人脸视频。Zhao等人（2022）考虑了全局信息流和局部信息流，以获取用于检测的全面而微妙的时空线索。Yu等人（2023）提出一种基于全局不一致视图和多时间尺度局部不一致视图的增强型多尺度时空不一致放大器，利用全局特征指导局部特征，以放大无法检测的时空异常。然而研究发现，基于CNN的视频检测方法倾向于处理局部特征，对全局时空特征或跨帧相关信息的捕捉不足，导致泛化效果有限。

1.2.2 基于Transformer的Deepfake检测方法

为了有效提取视频中的时间信息，研究人员提出多种机制。传统的RNN按顺序处理序列中的数据，但在处理长期依赖关系时存在局限性。为解决这个问题，引入了长短期记忆（LSTM）机制，它利用门控状态来控制信息的流动，更有效地保留长期时间信息，并过滤不重要的信息。此外，研究中还采用了Transformer机制，它能有效处理序列数据中的长期依赖关系。与LSTM相比，Transformer能更有效地捕捉远帧之间的关系，提供更好的时间关系建模。最近的研究将变换器扩展到计算机视觉（CV）任务，并衍生出一系列变体，如视觉Transformer（visual Transformer，ViT）。ViT直接将Transformer应用于图像块序列，以进行图像分类。还有人提出一些基于Transformer的端到端深度防伪检测方法。Wang等人（2022b）提出一种多模态多尺度Transformer，它对不同大小的斑块进行操作，以检测图像在不同空间域和频域的局部不一致性。Heo等人（2023）提出一种高效的视觉Transformer模型，以提取局部和全局

空间特征，用于深度伪造检测。Transformer还用于视频级的伪造检测，Yang等人(2023)利用视觉和听觉模式之间固有的同步模式，使用Transformer进行联合检测。Zhao等人(2023)提出可解释时空视频变换器，它由一种新的分解的时空自注意力和自减机制组成，用于捕获空间伪影和时间不一致性。Wang等人(2023)使用Timesformer和I3D(inflated 3D ConvNet)网络分析了整个面部和嘴唇区域的动态不一致性。

通过上述分析发现，对于视频级Deepfake检测，基于CNN的检测方法通常先提取空间特征，然后提取时间特征，并从全局角度识别局部伪造区域。同时，利用Transformer网络进行的Deepfake检测更侧重于捕捉全局的时空不一致性，这在一定程度上影响了对伪造信息的全面捕捉。因此，对模型泛化能力的提升是有限的。为了增强泛化能力，本文提出MBMD模型，该模型从频率、空间和时空域的多个分支捕捉伪造痕迹，并将这些不同域的全局和局部特征信息进行整合。

2 本文方法

2.1 动机

随着Deepfake技术的不断发展，传统的检测方法面临着越来越复杂的挑战。为了提高检测的有效性，研究者逐渐认识到单纯依赖某一信息域(如空间特征或时空特征)存在局限性，逐步将频域、空域与时域信息进行结合。然而，尽管现有的多模态方法在不同程度上利用了频率信息，这些方法仍存在一些明显问题。例如，很多方法未能进行选择性提取频率信息中的有用频率分量，而是处理了大量冗余信息。这导致噪声增加，影响了检测的准确性。此外，现有方法在结合局部和全局特征时存在局限，许多方法依赖单一网络结构进行特征提取，未能充分发挥CNN和Transformer等不同网络的优势。

具体而言，传统的Deepfake视频检测方法主要集中在时空特征的提取和分析。尽管许多现有方法，如Gu等人(2022a)、Zhu等人(2024)已经开始融合时空域、空间域和频域信息，但在频率信息的有效利用上仍面临一些挑战。尤其是在处理光照变化较大的复杂场景时，空间特征容易受到光照影响，而频率信息则能够提供更稳定的特征表示。因此，频域信息的有效利用成为提高检测精度的关键。然而，现有的基于频率的方法大多使用全频域信息，但未能有效选择性提取其中的有用信息。频率信息中的冗余和噪声未得到充分去除，导致检测效果受到影响。从图1可以观察到，无论是高质量还是低质量图像，RGB图像与低频图像在视觉上差别不显著，但在高频区域，真假人脸的差异更加明显。这一发现为频域信息在伪造检测中的应用提供了新的思路。低频信息和空间特征承载着图像的整体结构和轮廓，然而低频信息中的噪声可能会干扰模型对局部细节特征的学习。因此，现有方法未能有效处理低频信息中的冗余与噪声，导致模型在捕捉局部细节时受到干扰。为解决这一问题，本文提出通过去除频率信息中的低频分量，并保留高频分量，有效减少冗余信息和噪声，从而突出图像中的细节和伪造痕迹。

此外，现有的Deepfake视频检测方法通常依赖于卷积神经网络(CNN)来捕捉图像的全局和局部时空特征。尽管CNN在捕捉局部细节特征方面具有显著优势，但在捕捉帧序列的全局时空特征时仍然存在局限性，尤其在捕捉长时间跨度中的时空关系时，其效果会下降。另一方面，使用Transformer进行检测的方法，虽然能够很好地建模全局时空信息，但

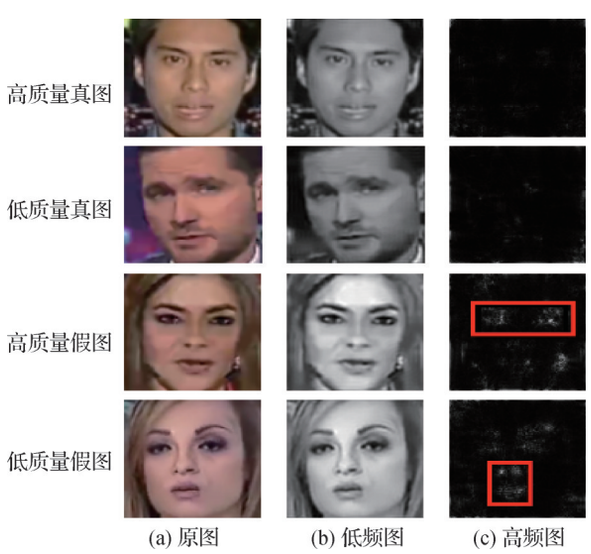


图1 不同质量的真假图像在RGB、低频和高频上的差异
(假图像的伪造区域用红框标记)

Fig. 1 Differences in RGB, low-frequency, and high-frequency domains for real and fake images of different quality (the forged regions in the fake images are marked with red boxes) ((a) original images; (b) low-frequency images; (c) high-frequency images)

对局部细节的关注不足，这限制了其对伪造痕迹的捕捉能力和泛化能力。针对现有方法的不足，本文提出一种结合视觉Transformer和CNN的方法，利用视觉Transformer处理帧序列来捕获全局时空信息，而CNN则专注于提取局部空间特征。这种结合能够充分利用CNN在局部特征提取中的优势，同时弥补Transformer在局部细节捕捉方面的不足，使模型可以更全面地捕捉时空信息，特别是在频率信息的补充下，模型能够进一步增强对伪造痕迹的感知能力。

综上所述，本文提出一种基于多域特征融合的多分支网络方法，通过结合视觉Transformer和CNN的优势，能够更有效地从全局和局部角度提取特征，并确保频率信息得到高效利用。通过整合频率、空间和时空域的信息，克服现有方法在频率信息选择性利用和全局与局部特征结合上的不足，从而提升Deepfake检测的泛化能力。

2.2 概述

拟议的网络框架如图2所示。具体而言，对于频率流分支，对帧序列中的图像进行DCT变换，去除引入额外噪声和信息冗余的低频成分，保留涉及微妙结构变化的高频成分，以获得更详细的频率信息；对于空间流分支，使用ResNet-34处理帧序列中的RGB图像，主要通过利用设计的空间特征增强块（SEB）增强多个尺度的浅层网络特征，以获得局部浅层特征（low shallow features，LSF），并使网络更加关注局部异常区域；对于时空流分支，使用视觉Transformer处理帧序列获得全局高层特征（global high-level features，GHF）以感知全局时空信息，通过设计的信息补充块（ISB），将空间流中LSF与GHF相结合，得到更全面的全局局部时空特征，使网络在保留全局时空信息的同时，能够更加关注与视频内容高度相关的局部异常区域。最后，通过设计的交互融合模块（ISB）将不同域的特征进行整合增强，从而综合利用频率、空间和时空信息进行Deepfake检测。

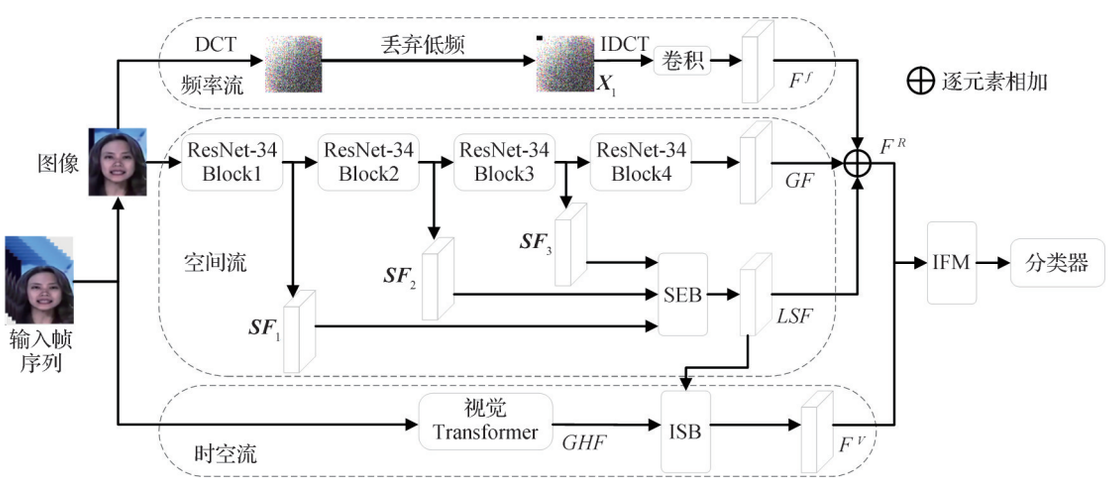


图2 所提网络架构

Fig. 2 Overview of the proposed network

2.3 频率流

Jia等人（2022）指出假人脸在高频段的能量比真人脸更丰富。在离散余弦变换（DCT）中，低频段的能量集中在左上角，高频段的能量集中在右下角。因此，对RGB图像进行DCT变换，丢弃频域图中的低频分量并保留高频分量，由此得到的频率图$\boldsymbol{X}_1$能够更有效地聚焦图像细节。对于一个$M \times N$的图像$\boldsymbol{f}$，其DCT变换$F(u,v)$可表示为

$$F = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x,y)\cos\left[\frac{\pi}{N}(x+\frac{1}{2})u\right]\times \cos\left[\frac{\pi}{M}(y+\frac{1}{2})v\right] \tag{1}$$

式中，$f(x,y)$表示原始图像在位置(x,y)处的像素值，$F(u,v)$表示DCT频域图像在位置(u,v)处的系数。$\alpha(u)$和$\alpha(v)$是归一化系数，可定义为

$$\alpha(u)=\begin{cases}\sqrt{\dfrac{1}{N}} & u=0\\ \sqrt{\dfrac{2}{N}} & u>0\end{cases} \tag{2}$$

$$\alpha(v)=\begin{cases}\sqrt{\dfrac{1}{M}} & v=0\\ \sqrt{\dfrac{2}{M}} & v>0\end{cases} \tag{3}$$

此外，去除频率信息中的低频分量的操作可表示为

$$F(u,v)_h=\begin{cases}0 & u^2+v^2<R^2\\ F(u,v) & u^2+v^2\geqslant R^2\end{cases} \tag{4}$$

式中，R是低频区域的半径，决定了低频分量的范围。

由于频率图不具有RGB图像的视觉特征，无法利用CNN进行特征提取，因此需要对频率图像进行逆DCT（inverse DCT，IDCT）处理，以重新获得RGB域的表示，最后，通过3×3卷积进行特征提取，从而获得频率特征（F^f）。逆DCT操作为

$$f(x,y)=\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\alpha(u)\alpha(v)F(u,v)\cos\left[\frac{\pi}{N}(x+\frac{1}{2})\right]u\cos\left[\frac{\pi}{M}(y+\frac{1}{2})v\right] \tag{5}$$

2.4 空间流

由于许多深度伪造检测方法直接在卷积神经网络的末尾层连接一个二元分类器，并基于全局特征进行决策。全局特征（global features，GF）能够帮助网络识别图像的整体一致性和自然性。例如，许多伪造方法可能修改图像的背景，或使伪造区域与背景之间的过渡不自然。GF能够捕捉到这些全局不一致的部分，从而有效识别伪造的痕迹。然而，这种方法往往忽视了隐藏在图像浅层特征中的伪造痕迹，因为某些伪造方法可能只篡改图像的特定局部区域。

为了解决这一问题，在空间流中提出一个空间特征增强块（SEB），如图3所示，旨在利用卷积神经网络从不同尺度上发现并提取空间伪造痕迹中更具鉴别性的局部浅层特征。

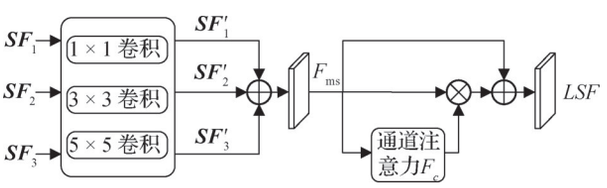


图3　空间特征增强块

Fig. 3　Structure of the spatial feature enhancement block

在空间特征增强块中，同时利用3个具有不同核大小的卷积层。具体而言，将从ResNet-34网络的浅层块中提取到的多个特征图（$\boldsymbol{SF}_1$，$\boldsymbol{SF}_2$，$\boldsymbol{SF}_3$）作为输入，同时送入多尺度卷积层，如图4所示。这些卷积层分别使用内核大小为1、3、5的卷积处理，生成对应的多尺度特征图（$\boldsymbol{SF}'_1$，$\boldsymbol{SF}'_1$，$\boldsymbol{SF}'_3$）。通过不同尺度的卷积操作，能够捕获不同大小的特征模式，从而获得更丰富、更全面的局部特征表示。随后，将这些输出的多尺度特征图连接以形成多尺度浅层特征。这些多尺度浅层特征汇集了不同层次信息，能够增强网络对图像细节的感知能力。具体而言，这一过程可以描述为

$$\boldsymbol{SF}'_1=conv_{1\times1}(\boldsymbol{SF}_1)\oplus conv_{3\times3}(\boldsymbol{SF}_1)\oplus conv_{5\times5}(\boldsymbol{SF}_1) \tag{6}$$

$$\boldsymbol{SF}'_2=conv_{1\times1}(\boldsymbol{SF}_2)\oplus conv_{3\times3}(\boldsymbol{SF}_2)\oplus conv_{5\times5}(\boldsymbol{SF}_2) \tag{7}$$

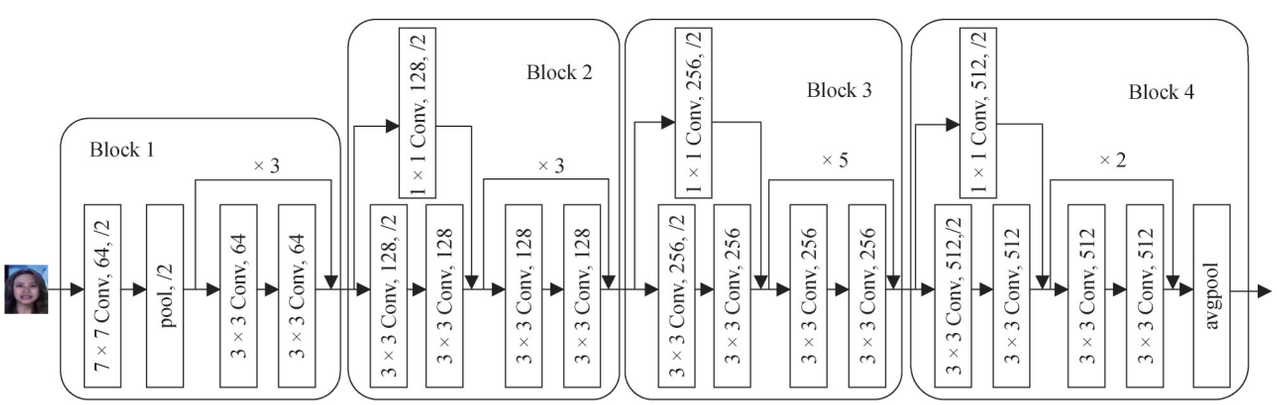


图4　ResNet-34网络架构

Fig. 4　Structure of the ResNet-34

$$\boldsymbol{SF}_3' = conv_{1\times1}(\boldsymbol{SF}_3) \oplus conv_{3\times3}(\boldsymbol{SF}_3) \oplus conv_{5\times5}(\boldsymbol{SF}_3) \tag{8}$$

$$\boldsymbol{F}_{ms} = \sum_{i=1}^{3} \boldsymbol{SF}_i' \tag{9}$$

式中，$\boldsymbol{SF}_i'(i = 1, 2, 3)$表示对特征图($\boldsymbol{SF}_i$)进行不同尺度卷积后得到的特征，⊕表示逐元素相加。

考虑到多尺度特征增强主要集中在空间维度信息的处理上，通过引入通道注意力块(Hu等，2018)加强特征图在通道维度上的关联，如图5所示。

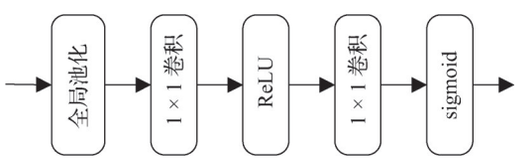


图5　通道注意力块

Fig. 5　Structure of the channel attention block

通道注意力块能够综合考虑所有通道的信息并确定每个通道的重要性，增强了特征在深度维度上的表征能力，从而提高模型对于关键特征的感知和利用。首先，将通道注意力块与多尺度浅层特征相乘，这有助于强调通道间的重要性，即增强模型对于特定通道的关注度。然后，将通道注意力块的输出与原始的多尺度浅层特征相连接，获得局部浅层特征(LSF)，具体为

$$\boldsymbol{LSF} = \boldsymbol{F}_{ms} \oplus (\boldsymbol{F}_{ms} \otimes \boldsymbol{F}_c) \tag{10}$$

式中，⊗表示逐元素相乘，$\boldsymbol{F}_c$表示通道注意力块。这一特征整合了不同尺度和通道的信息，使得模型更全面地理解图像中的结构和细节，进而加强对局部异常区域的关注。

此外，GF是通过ResNet-34网络的末尾层提取的图像全局信息。此部分特征通常包括图像的大致结构、全局纹理信息和宏观背景特征。与LSF不同，全局特征侧重于整个图像的空间信息，帮助模型理解图像的整体一致性与背景结构，尤其当伪造涉及大范围的背景修改或全局结构变化时，GF提供了重要的上下文信息，确保网络能够识别这些大范围的伪造痕迹。通过将LSF与GF相结合，模型能够在全局特征的基础上补充局部信息，从而更好地识别细节。这种全局与局部特征的互补不仅增强了网络对伪造图像细节的识别能力，同时也保留了全局一致性的判断能力。最终，网络在检测过程中能够同时关注图像的全局信息(如背景)和局部信息(如面部或纹理)。

2.5　时空流

由于视觉Transformer(ViT)网络更擅长捕捉图像中的全局时空特征，通过自注意力机制聚焦图像的全局信息，捕捉长距离的时空依赖关系，从而理解视频中的整体结构和时空动态。考虑到伪造痕迹会出现在图像的局部区域(例如面部特征区域或细节纹理)，ViT可能会忽视这些细微的局部变化或伪造痕迹，导致局部细节的丢失或不够敏感。为了解决ViT网络在关注全局时空信息的同时，难以有效捕捉到局部细微异常的问题，本文设计了一个信息补充块(ISB)，通过建立从一个域到另一个域的信息流来增强后者的特征表示。该设计优化了局部和全局时空特征的整合，使得模型能够更全面地关注伪造区域，特别是局部异常区域，而不是仅仅依赖全局时空信息。

利用从卷积神经网络捕获的局部空间特征来对视觉Transformer网络提取的全局高层特征进行补充增强。以这种方式，网络能够同时关注视频中的局部细节和全局结构，更全面地捕获全局与局部区域中出现的时空伪造痕迹。具体过程如图6所示。

具体而言，将来自CNN的局部浅层特征(LSF)通过最大池化、卷积、sigmoid函数处理后，再与原始的局部浅层特征(LSF)相乘，获得局部关键特征。这一过程旨在增强LSF中各个通道的相关性，从而使网络更好地聚焦于最具辨别力的特征。其中，最大池化用于减少特征图的空间维度并保留最显著的特征，这有助于网络突出关键特征区域并增强网络对局部特征的表征能力。卷积操作进一步处理池化后的特征图，提取特征间的高层次信息。sigmoid函数对每个通道的重要性进行归一化处理，强调每个通道的重要性。这一系列操作增强了局部关键特征，同时抑制了不重要的局部特征。操作过程可表

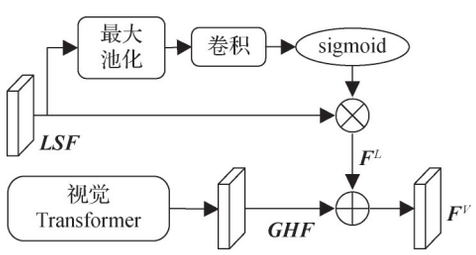


图6　信息补充块

Fig. 6　Structure of the information supplement block

示为

$$\boldsymbol{F}^{L}=\boldsymbol{LSF}\otimes\sigma(conv(MaxPool(\boldsymbol{LSF}))) \quad (11)$$

式中，σ 是sigmoid函数。

经过上述处理后，将强化后的局部关键特征作用于视觉Transformer网络获得的全局高层特征（GHF）中，使网络能够同时关注全局和局部的时空不一致性，从而帮助网络更加准确地识别帧序列中的重要特征。获得的时空特征可表示为

$$\boldsymbol{F}^{V}=\boldsymbol{GHF}\oplus\boldsymbol{F}^{L} \quad (12)$$

ISB通过结合CNN提取的局部空间特征和ViT提取的全局时空特征，能在保留全局时空信息的同时，提升对局部伪造痕迹的感知能力。这一过程不仅增强了网络对细节和异常区域的关注，还将局部和全局信息整合到整体特征表达中，使获得的时空特征更加全面和细致。

2.6 交互融合

由于帧序列中的单帧图像和序列图像包含不同的信息，在未经信息交流的情况下直接连接，会丢失重要的伪造线索。因此，本文设计了一个交互融合模块（IFM）来实现不同域之间信息的相互促进，如图7所示。

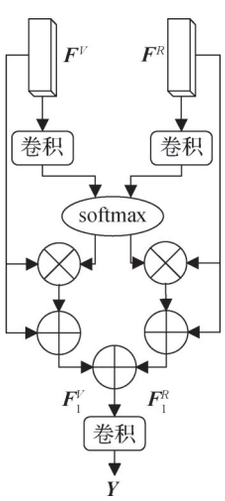


图7 交互融合模块

Fig. 7 Structure of the interactive fusion module

具体而言，对于从图像中获得的频率和空间特征 $\boldsymbol{F}^{R}$ 和从帧序列中获得的时空特征 $\boldsymbol{F}^{V}$，它们表示了来自不同领域的特征，通常具有不一致的维度。卷积操作可将其映射到相同的维度空间。以时空特征为例，通过softmax函数对 $\boldsymbol{F}^{R}$ 进行归一化处理，再与 $\boldsymbol{F}^{V}$ 进行相乘，这一过程利用频率和空间特征突出了时空域中某些通道的特征，增强了对特定时空特征的关注。然后将处理后的特征与原始时空特征相结合，得到增强的时空特征（$\boldsymbol{F}_1^{V}$）。$\boldsymbol{F}_1^{R}$ 的处理过程与此类似。最终，将增强后的特征连接，作为分类器的输入。这可以概括为

$$\boldsymbol{F}_1^{R}=\boldsymbol{F}^{R}\oplus\boldsymbol{F}^{R}\otimes softmax(conv(\boldsymbol{F}^{V})) \quad (13)$$

$$\boldsymbol{F}_1^{V}=\boldsymbol{F}^{V}\oplus\boldsymbol{F}^{V}\otimes softmax(conv(\boldsymbol{F}^{R})) \quad (14)$$

$$\boldsymbol{Y}=conv(concat(\boldsymbol{F}_1^{R},\boldsymbol{F}_1^{V})) \quad (15)$$

式中，$\boldsymbol{F}_1^{R}$ 是增强的频率空间特征，$\boldsymbol{F}_1^{V}$ 是增强的时空特征。

在交互增强操作中，通过将网络中一个领域学习到的特征共享到另一个领域，促进了不同域特征之间的信息传递和交互学习。这种操作利用不同域的特征信息来建立它们之间的联系，从而丰富特征表达。通过交互融合模块（IFM），不同域的特征得到了相应的增强，提高了原始特征的多样性和丰富性，从而增强了特征的表达和辨别能力。最终输入分类器的特征 Y 是综合利用不同域特征的结果，能够更好地提高模型的泛化能力。最后，使用交叉熵损失来训练网络，从而获得最终的检测结果。

3 实验结果与分析

3.1 实验设置

在实验中，使用ResNet-34（He等，2024）和视觉Transformer（Dosovitskiy等，2020）作为拟议网络的骨干网络。对于每个数据集，使用人脸检测器Dlib（King，2009）提取人脸，并将人脸对齐为 224 × 224 像素。在训练阶段，采用Adam优化器（Kingma和Ba，2015）对二元交叉熵损失进行优化，批量大小为16，初始学习率设置为0.000 1。所有数据集的周期总数设为40。该任务基于PyTorch实现，并在NVIDIA GTX 3090 GPU上进行训练。

3.2 实验数据集

本文使用FaceForensics++（Rossler等，2019）、Celeb-DF-v2（Li等，2020b）和DFDC（DeepFake Detection Challenge）（Dolhansky等，2020）3个数据集进行实验。

FaceForensics++是一个常用的深度伪造基准数据集，涵盖了不同级别的视频质量，包括高质量（high quality，HQ）和低质量（low quality，LQ）。每个视频数据集由1 000个原始视频和4 000个使用

4种不同技术生成的篡改视频组成：DF（deepfakes）、F2F（face to face）、FS（faceswap）和NT（neuraltextures）。

Celeb-DF-v2包含590个真实视频和5 639个篡改视频。真实视频来自公开的YouTube视频剪辑。

DFDC由Facebook DeepFake Detection Challenge发布，包含约5 000个采用了多种未知处理方法的视频。这些视频中的人脸可能经过部分篡改，导致即使是最先进的检测器在这个具有挑战性的DFDC数据集上也表现不佳。

3.3 评估指标

在实验中，采用准确度得分（accuracy，ACC）和曲线下面积（area under curve，AUC）作为评价指标。AUC表示可分性的程度或量度，它通过区分不同类别来评估模型准确划分不同区域的能力，AUC越高，说明模型越好。

3.4 数据集内评估

数据集内测试分别对FF++、Celeb-DF-v2和DFDC数据集进行。比较结果如表1所示。总体而言，MBMD在3个数据集上取得了相当出色的结果。不过，在Celeb-DF-v2数据集中的实验准确率略低。这可能是由于Celeb-DF-v2数据集中真假样本的比例较大，可能导致特征表征学习不足。

相比之下，MBMD在FF++数据集上的AUC值分别比HolisticDFD和DIIR（deepfake detection via inter-frame inconsistency recomposition）高5.39%和4.39%。此外，在DFDC数据集上，MBMD的ACC值分别比DIIR和ISTVT（interpretable spatial-temporal video Transformer）高出1.54%和1.83%。由于FF++数据集相较于其他数据集（如Celeb-DF-v2）具有较为复杂的背景、光照变化以及伪造痕迹的分布，特别是在光照变化和阴影干扰较大的场景下，伪造的痕迹更加难以被捕捉。所提方案通过结合频率特征、空间特征和时空特征，克服了光照一致性问题，使得网络能够从不同域获得更全面的伪造痕迹，从而更好地应对FF++数据集的复杂场景。

据观察，ISTVT和HolisticDFD主要通过提取时空特征来检测伪造，没有利用额外的频率信息。DIIR主要利用相邻帧的光照不一致性进行伪造检测。然而，该方案中的图像分解模块（image decomposition module，IDM）存在一定局限性，特别是在一些复杂区域，如头发等细节区域，光照分解的效果并不理想。在其他数据集尤其是Celeb-DF-v2数据集中的图像分辨率较高且光照分布均匀，这使得DIIR在此数据集上的表现更为优秀，因为其光照分解方法能够有效地处理光照较为均匀的图像。

表1 数据集内检测性能评估

Table 1 The intra-dataset detection performance evaluation

/%

方法	Params/M	FF++		Celeb-DF-v2		DFDC	
		ACC	AUC	ACC	AUC	ACC	AUC
Two-branch（Masi等，2020）	67.1	91.38	95.11	–	76.65	–	–
S-MIL（Li等，2020a）	42.2	95.51	97.45	97.12	98.29	91.67	93.13
TD-3DCNN（Zhang等，2021）	–	79.09	72.22	81.08	88.83	82.64	88.97
DIANet（Hu等，2021）	32.7	95.57	97.73	97.74	98.41	92.19	94.24
LipForensics（Haliassos等，2021）	36.0	96.50	98.90	97.10	99.10	–	–
STI（Gu等，2021）	32.5	96.69	97.59	97.16	**99.78**	91.17	89.80
DDN（Pu等，2022）	–	95.60	<u>99.00</u>	96.50	98.90	–	–
Finfer（Hu等，2022a）	–	–	93.91	90.47	93.30	80.39	82.88
ISTVT（Zhao等，2023）	–	**97.57**	–	**99.80**	–	92.10	–
HolisticDFD（Anas Raza等，2023）	11.5	–	94.15	–	96.24	–	92.60
DIIR（Zhu等，2024）		90.31	95.15	<u>99.11</u>	<u>99.74</u>	<u>92.39</u>	**97.37**
MBMD（本文）	131.0	<u>96.97</u>	**99.54**	94.80	97.77	**93.93**	<u>97.31</u>

注：加粗、下划线字体分别表示各列最优、次优结果。“–”表示无结果。

然而，所提方案通过提取帧序列来拟合更复杂的现实场景，采用了非连续帧序列的输入方式。这一方法能够有效处理一些复杂的伪造痕迹，特别是在光照变化或背景复杂的情况下，增强了模型对伪造区域的感知。然而，由于采用非连续帧序列，部分伪造线索无法在帧间连续捕捉，导致检测效果有所局限。虽然这种方法在某些复杂场景下能够提供更强的伪造识别能力，但它也未能充分利用连续帧之间的光照一致性。通过对比参数量发现，使用单一网络进行检测的方案（如HolisticDFD）具有较少的参数量，特别是HolisticDFD采用了轻量化网络架构，因此其参数量较少。而所提方法结合了CNN和视觉Transformer网络，这种组合提供了更强的特征表达能力，但也相应增加了计算量和存储需求。上述结果表明，尽管所提方法相比于其他方法具有更高的计算开销，但在性能提升上展现了明显的优势，尤其是在处理复杂场景时。

3.5 跨数据集评估

随着伪造技术的不断创新和发展，伪造检测模型需要更具灵活性和适应性，以有效地识别和对抗不断涌现的伪造内容。为了评估所提出的网络在未知数据集上的泛化能力，本文在FF++上对其进行训练，并在未知的Celeb-DF-v2和DFDC数据集上进行测试。结果如表2所示。

从表2可以观察到，与数据集内的结果相比，跨数据集实验结果普遍呈现出不同程度的下降，尽管如此，拟议的MBMD在不同数据集之间仍然表现出色，这证明了其卓越的泛化能力。就AUC而言，在未知的Celeb-DF-v2数据集上，与ECGL、ISTVT和DIIR相比，拟议的MBMD分别提高3.04%、4.8%和13.65%。同样，在未知的DFDC数据集上，比ISTVT和ECGL分别提高4.43%和5.54%。对于ECGL，它利用空间和时间模块顺序处理全局和局部信息，并设计了全局流引导损失，利用全局信息引导局部信息的选择。然而，有些伪造方法只篡改面部细微区域，从全局角度检测的效果并不明显。ISTVT通过空间注意力关注面部特征的混合边缘和异常边界，并利用时间注意力解决短期帧不一致问题。相比之下，所提出的MBMD不仅利用频率信息，还整合了局部空间和全局时空信息，有利于捕捉伪造线索的不同方面，从而实现了更好的泛化性能。

表2 未知数据集泛化性能

Table 2 Generalization performance on unseen datasets

/%

方法	FF++上训练			
	Celeb-DF-v2		DFDC	
	ACC	AUC	ACC	AUC
Two-branch（Masi等，2020）	–	73.41	–	–
S-MIL（Li等，2020a）	73.29	75.86	64.74	66.79
TD-3DCNN（Zhang等，2021）	66.02	57.32	**82.21**	55.02
DIANet（Hu等，2021）	70.83	73.94	65.95	68.65
LipForensics（Haliassos等，2021）	–	82.40	–	73.50
STI（Gu等，2021）	74.67	76.17	65.95	67.22
DIL（Gu等，2022b）	81.18	82.82	70.93	72.52
ECGL（Zhao等，2022）	<u>83.17</u>	<u>85.86</u>	71.39	73.09
Finfer（Hu等，2022a）	–	70.60	69.45	70.39
ISTVT（Zhao等，2023）	–	84.10	–	<u>74.20</u>
DIIR（Zhu等，2024）	–	75.25	–	–
MBMD(本文)	**85.80**	**88.90**	<u>77.20</u>	**78.63**

注：加粗、下划线字体分别表示各列最优、次优结果。“–”表示无结果。

3.6 频域变换性能分析

为了确保方法的合理性和适用性，本文借鉴了Qian等人（2020）提出的F3-Net方案中的提取方法。该方案在频域信息利用方面表现优异，且具有较强的代表性。基于F3-Net的分析，选择将半径R设置为0.125，该值在保证图像特征提取效果的同时，能够有效去除低频信息，显著提升处理效率。此外，对离散余弦正逆变换（DCT和IDCT）的处理时间和内存消耗进行了详细计算。实验结果如表3所示。

从表3可以观察到，IDCT变换所需的时间不到DCT变换的两倍，且整体处理时间仍保持较短。尽

表3 正逆变换计算负担（时间与内存）分析

Table 3 Analysis of computational burden (time and memory) of forward and inverse transforms

数据集	帧序列/幅	总时间/s		内存消耗/MB
		DCT	IDCT	
FF++	182 522	86.54	156.51	26 201
Celeb-DF-v2	127 994	60.49	108.91	18 374
DFDC	11 257	5.24	9.57	1 616

管处理的帧序列数量庞大(达到数十万幅),导致内存消耗较高,但正逆变换带来的处理速度提升使得该方法在大规模视频帧序列处理中的应用依然具备显著优势。这种方法能够在保证图像质量的同时,进一步优化处理效率,展现出其在大规模数据处理中的卓越性能。

3.7 消融研究

为了验证所提网络的每个组成部分对提高泛化能力的影响,本文在FF++上对模型进行训练,然后在FF++以及未见过的Celeb-DF-v2和DFDC数据集上进行了测试。

3.7.1 频率分量的影响

为了研究频率信息对模型泛化性能的影响,本文进行了几组实验,结果如表4所示。从跨数据集实验中可以观察到,包含所有频率信息的实验模型在性能上低于仅利用时空特征或时空特征与空间特征结合的模型。这一结果可能是由于未对频率信息进行选择性利用,导致模型学习到了过多的冗余信息。在不同场景中,频域信息包含了许多不重要的细节,甚至干扰信息,这些冗余信息的存在反而影响了模型的特征学习,导致了判别的失误。

表4 频率分量对泛化能力的影响

Table 4 Effect of frequency components on generalization ability

/%

模型	FF++上训练		
	FF++	Celeb-DF-v2	DFDC
时空	90.63	80.43	66.24
空间+时空	93.76	82.40	69.30
空间+时空+频率	92.85	80.11	62.50
空间+时空+低频	90.66	81.78	66.58
空间+时空+高频	**96.97**	**85.80**	**77.20**

注:加粗字体为各列最优值。

相比之下,利用高频信息的模型表现最好。在Celeb-DF-v2和DFDC数据集上,相比于使用所有频率信息的模型,利用高频信息的模型分别提高5.69%和14.7%;而与仅利用低频信息的模型相比,分别提高4.02%和10.62%。这种差异表明,低频成分包含了大量的背景信息和相对不重要的图像细节,对伪造视频的区分作用有限。仅利用低频信息不仅无法有效识别伪造视频中的关键特征,还可能引入噪声和冗余信息,进而降低模型的泛化能力。本文方法通过去除低频成分,有效减轻了信息干扰,提高模型对输入数据的感知能力。保留的高频成分则主要包含了图像中的重要细节,如纹理、边缘等,这使得模型能够更加聚焦于图像的关键特征。这一优势有助于模型更好地识别视频内容中的细微变化和结构特征,从而显著提升了模型的泛化性能。

3.7.2 局部空间特征和全局时空特征的影响

为了研究局部空间特征和全局时空特征对模型泛化能力的影响,本文在多个场景进行实验,以验证不同模块对模型性能的具体作用。实验结果如表5所示。从结果可以观察到,模型3的总体表现最低,可能是因为该模型在空间流中没有利用局部特征,而时空流则过于侧重全局时空信息,对局部异常的关注不足,导致模型捕获的特征信息偏向于全局,忽略了局部异常区域的细节,这使得模型难以有效识别局部伪造痕迹,表现较差。模型4的性能相比模型1和模型3有所提升,但仍不如模型2。这可能是由于模型4仅将不同的局部浅层特征与全局高层特征进行简单相加,缺乏进一步的信息处理。尤其是在没有通过空间特征增强块(SEB)的情况下,局部特征未经过多尺度增强和注意力机制的优化,导致模型无法有效捕捉局部关键特征,从而影响了对全局和局部时空信息的综合捕获。相比之下,模型2通过引入空间特征增强块(SEB),能够更好地捕捉图像中的丰富特征。SEB通过对不同尺度特征进行增强并优化网络对与任务相关的特征通道的关注,抑制与任务无关的噪声通道,从而提升了模型的特征学习能力。这种方法有效增强了模型的泛化能力,使得模型在不同的伪造视频场景中能够保持较高的检测性能。此外,模型6的表现优于模型2和模型5,这一改善可能是因为模型2虽然增强了局部空间信息,但没有有效地将局部特征与全局高层特征结合,造成时空流中的特征无法充分捕捉全局与局部的时空异常。模型5则试图通过信息补充块(ISB)弥补这一缺陷,然而,由于缺乏SEB的多尺度增强和特征挖掘,模型未能有效捕捉局部异常,导致其性能仍未达到理想水平。模型7展现了最优的性能,相比于模型2和模型5仅使用网络的单一模块,本文使用的模型7通过SEB捕捉帧序列中的局部关键异常区域,并结合ISB弥补时空流在捕捉全局时

空信息时的局限性。这种综合方法使得模型能够同时关注视频中的局部细节和全局结构，更准确地识别视频中的关键特征，从而显著提高模型的泛化性能。

表 5　局部/全局信息对泛化能力的影响

Table 5　Effect of local-global information on generalization ability

/%

模型	$SF_1+SF_2+SF_3$	空间特征增强块	全局高层特征	信息补充块	FF++上训练		
					FF++	Celeb-DF-v2	DFDC
模型 1	√	–	–	–	88.48	80.56	65.16
模型 2	–	√	–	–	94.48	83.40	73.60
模型 3	–	–	√	–	87.97	78.20	63.05
模型 4	√	–	√	–	91.46	81.80	66.11
模型 5	√	–	–	√	93.82	81.42	66.80
模型 6	–	√	√	–	95.52	83.23	73.95
模型 7	–	√	–	√	**96.97**	**85.80**	**77.20**

注：加粗字体表示各列最优结果。“√”和“–”分别表示使用和未使用对应模块。

3.7.3　交互融合模块的影响

实验分析了交互融合模块的影响，结果如表 6 所示。可以发现，在 FF++、Celeb-DF-v2 和 DFDC 数据集上，使用交互融合模块（IFM）的结果比直接连接图像特征和帧序列特征的结果分别提高 2.39%、3% 和 4.6%。这可能归功于该模块引入的跨域信息交互机制。通过对不同领域的特征进行交叉处理，建立起不同领域特征之间的关系，使其相互影响，丰富特征表达。通过交互式特征增强操作，重要的特征得到增强，而不重要的特征则被削弱，这有利于提取出更具区分度和代表性的特征，从而提高网络的泛化能力。相比之下，直接连接图像特征无法有效利用不同领域的信息，可能会导致关键信息的丢失，从而降低泛化性能。实验结果验证了交互融合模块的有效性。

表 6　交互融合模块对泛化能力的影响

Table 6　Effect of the interactive fusion module（IFM）on generalization ability

/%

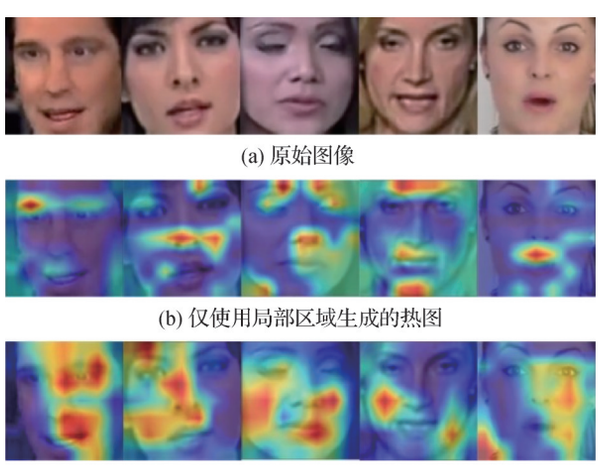
模型	FF++上训练		
	FF++	Celeb-DF-v2	DFDC
F^R+F^V	94.58	82.8	72.6
交互融合模块	**96.97**	**85.8**	**77.2**

注：加粗字体表示各列最优结果。

3.8　可视化

为进一步说明拟议网络的优势，本文随机选取了不同样本进行可视化分析，如图 8 所示。图 8（b）显示了仅使用局部区域生成的热图（Selvaraju 等，2020），而图 8（c）显示了本文方法生成的热图。可以观察到，依赖局部区域的模型无法全面分析伪造情况，而本文方法则利用多种伪造线索进行更精确的检测。

此外，图 9 的可视化结果进一步验证了拟议网络中局部和全局信息的优势。可以发现，无 SEB（空



(a) 原始图像

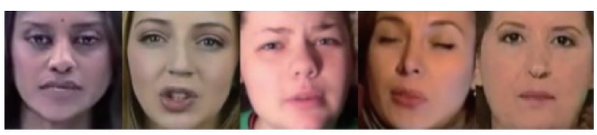
(b) 仅使用局部区域生成的热图

(c) 提出的方法生成的热图

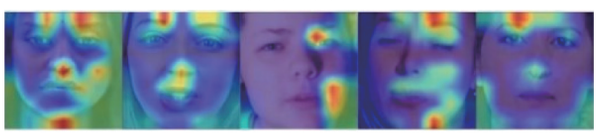
图 8　不同样本的可视化

Fig. 8　Visualization of different samples ((a) original images; (b) the generated heatmaps by using only local regions; (c) the generated heatmaps of the proposed method)

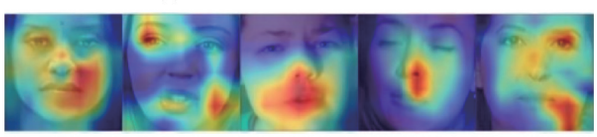
间增强块)和ISB(信息补充块)的模型只关注较为分散的区域,导致对伪造信息的捕捉不足。拟议网络通过同时使用SEB和ISB来捕捉伪造线索,重点关注更全面的区域,包括人脸的中心区域和边缘区域。这表明,使用两个模块可以让网络同时关注局部和全局异常区域,从而捕捉到更全面的伪造线索,提高模型的泛化能力。



(a) 原始图像



(b) 无空间增强块和信息补充块



(c) 有空间增强块和信息补充块

图9 有无SEB + ISB的可视化

Fig. 9 Visualization without SEB + ISB and with SEB + ISB ((a) original images; (b) visualization without SEB + ISB; (c) visualization with SEB + ISB)

4 结 论

本文提出一种基于多域特征融合(MBMD)的多分支网络,综合利用频域、空间域和时空域信息,更全面地捕捉伪造线索,从而增强Deepfake检测的泛化性。为此,需要从帧序列的不同域中挖掘伪造痕迹。在频率流中,对图像进行DCT变换,去除引入额外噪声和冗余信息的低频成分,保留高频成分,以捕捉图像中细微结构变化的频率特征。在空间流中,设计了一个空间特征增强块(SEB),对CNN的浅层特征进行多尺度增强。生成的局部浅层特征(LSF)主要针对局部异常区域,能有效捕捉图像中的局部伪造痕迹。同时,在时空流中,设计了一个信息补充块(ISB),将局部浅层特征与视觉Transformer捕获的全局高层特征相结合,从而获得全面的全局—局部时空信息。由此,网络就能全面捕捉时空不一致性。最后,利用交互融合模块(IFM)增强和融合来自频域、空间域和时空域的信息,从而获得更全面、更详细的特征,用于Deepfake检测。在不同数据集上进行的实验表明,与一些最先进的方法相比,所提出的方法能够在未知数据集上表现出更强的泛化能力。然而,本文中帧序列的输入采用的是间隔采样方法,对连续帧的伪造检测还未深入探讨。在未来的研究中,将进一步研究连续帧中伪造检测的泛化性能和轻量化的设计,以及各种模态下的伪造检测技术,如音频和视频的结合。

参考文献(References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Amerini I, Galteri L, Caldelli R and Del Bimbo A. 2019. Deepfake video detection through optical flow based CNN//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE: 1205-1207 [DOI: 10.1109/ICCVW.2019.00152]

Anas Raza M, Malik K M and Ul Haq I. 2023. HolisticDFD: infusing spatiotemporal transformer embeddings for deepfake detection. Information Sciences, 645: #119352 [DOI: 10.1016/j.ins.2023.119352]

Bansal A, Ma S G, Ramanan D and Sheikh Y. 2018. Recycle-GAN: unsupervised video retargeting//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 122-138 [DOI: 10.1007/978-3-030-01228-1_8]

Chen B J, Li T M and Ding W P. 2022. Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM. Information Sciences, 601: 58-70 [DOI: 10.1016/j.ins.2022.04.014]

Chen H, Li Y Z, Lin D D, Li B and Wu J Q. 2023a. Watching the BiG artifacts: exposing DeepFake videos via Bi-granularity artifacts. Pattern Recognition, 135: #109179 [DOI: 10.1016/j.patcog.2022.109179]

Chen S, Yao T P, Chen Y, Ding S H, Li J L and Ji R R. 2021. Local relation learning for face forgery detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l]: AAAI Press: 1081-1088 [DOI: 10.1609/aaai.v35i2.16193]

Chintha A, Thai B, Sohrawardi S J, Bhatt K, Hickerson A, Wright M, et al. 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE Journal of Selected Topics in Signal Processing, 14 (5): 1024-1037 [DOI: 10.1109/JSTSP.2020.2999185]

Coccomini D A, Messina N, Gennaro C and Falchi F. 2022. Combining EfficientNet and vision transformers for video deepfake detection//

Proceedings of the 21st International Conference on Image Analysis and Processing. Lecce, Italy: Springer: 219-229 [DOI: 10.1007/978-3-031-06433-3_19]

Ding F, Kuang R S, Zhou Y, Sun L, Zhu X G and Zhu G P. 2024. A survey of Deepfake and related digital forensics. Journal of Image and Graphics, 29(2): 295-317 (丁峰, 匡仁盛, 周越, 孙珑, 朱小刚, 朱国普. 2024. 深度伪造及其取证技术综述. 中国图象图形学报, 29(2): 295-317) [DOI: 10.11834/jig.230088]

Dolhansky B, Bitton J, Pflaum B, Lu J K, Howes R, Wang M L, et al. 2020. The DeepFake detection challenge (DFDC) dataset [EB/OL]. [2024-11-05]. https://arxiv.org/pdf/2006.07397.pdf

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net: 1-21

Durall R, Keuper M, Pfreundt F J and Keuper J. 2019. Unmasking DeepFakes with simple features [EB/OL]. [2024-11-05]. https://arxiv.org/pdf/1911.00686.pdf

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D and Holz T. 2020. Leveraging frequency analysis for deep fake image recognition//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: PMLR: 3247-3258

Ganguly S, Ganguly A, Mohiuddin S, Malakar S and Sarkar R. 2022. ViXNet: vision Transformer with Xception Network for deepfakes based video and image forgery detection. Expert Systems with Applications, 210: #118423 [DOI: 10.1016/j.eswa.2022.118423]

Ganiyusufoglu I, Ngô L M, Savov N, Karaoglu S and Gevers T. 2020. Spatio-temporal features for generalized detection of deepfake videos [EB/OL]. [2024-11-05]. https://arxiv.org/pdf/2010.11844.pdf

Gu Q Q, Chen S, Yao T P, Chen Y, Ding S H and Yi R. 2022a. Exploiting fine-grained face forgery clues via progressive enhancement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI Press: 735-743 [DOI: 10.1609/aaai.v36i1.19954]

Gu Z H, Chen Y, Yao T P, Ding S H, Li J L, Huang F Y, et al. 2021. Spatiotemporal inconsistency learning for DeepFake video detection//Proceedings of the 29th ACM International Conference on Multimedia. [s. l.]: ACM: 3473-3481 [DOI: 10.1145/3474085.3475508]

Gu Z H, Chen Y, Yao T P, Ding S H, Li J L and Ma L Z. 2022b. Delving into the local: dynamic inconsistency learning for DeepFake video detection//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 744-752 [DOI: 10.1609/aaai.v36i1.19955]

Güera D and Delp E J. 2018. Deepfake video detection using recurrent neural networks//Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand: IEEE: 1-6 [DOI: 10.1109/AVSS.2018.8639163]

Haliassos A, Vougioukas K, Petridis S and Pantic M. 2021. Lips don't lie: a generalisable and robust approach to face forgery detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5037-5047 [DOI: 10.1109/CVPR46437.2021.00500]

He Q L, Peng C L, Liu D C, Wang N N and Gao X B. 2024. GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning. Neural Networks, 180: #106636 [DOI: 10.1016/j.neunet.2024.106636]

Heo Y J, Yeo W H and Kim B G. 2023. DeepFake detection algorithm based on improved vision transformer. Applied Intelligence, 53(7): 7512-7527 [DOI: 10.1007/s10489-022-03867-9]

Hu J, Liao X, Liang J W, Zhou W B and Qin Z. 2022a. FInfer: frame inference-based Deepfake detection for high-visual-quality videos//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 951-959 [DOI: 10.1609/aaai.v36i1.19978]

Hu J, Liao X, Wang W and Qin Z. 2022b. Detecting compressed Deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 32(3): 1089-1102 [DOI: 10.1109/TCSVT.2021.3074259]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Hu Z H, Xie H T, Wang Y X, Li J H, Wang Z Y and Zhang Y D. 2021. Dynamic inconsistency-aware DeepFake video detection//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI: 736-742 [DOI: 10.24963/ijcai.2021/102]

Ismail A, Elpeltagy M, Zaki M S and Eldahshan K. 2022. An integrated spatiotemporal-based methodology for deepfake detection. Neural Computing and Applications, 34(24): 21777-21791 [DOI: 10.1007/s00521-022-07633-3]

Jia S, Ma C, Yao T P, Yin B J, Ding S H and Yang X K. 2022. Exploring frequency adversarial attacks for face forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4093-4102 [DOI: 10.1109/CVPR52688.2022.00407]

Kim T, Cha M, Kim H, Lee J K and Kim J. 2017. Learning to discover cross-domain relations with generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 1857-1865

King D E. 2009. Dlib-ml: a machine learning toolkit. The Journal of Machine Learning Research, 10: 1755-1758

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-15

Li L Z, Bao J M, Yang H, Chen D and Wen F. 2019. FaceShifter: towards high fidelity and occlusion aware face swapping [EB/OL]. [2024-11-05]. https://arxiv.org/pdf/1912.13457.pdf

Li X D, Lang Y N, Chen Y F, Mao X F, He Y, Wang S H, et al. 2020a. Sharp multiple instance learning for deepfake video detection//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1864-1872 [DOI: 10.1145/3394171.3414034]

Li Y, Bian S, Wang C T, Polat K, Alhudhaif A and Alenezi F. 2023. Exposing low-quality deepfake videos of social network service using spatial restored detection framework. Expert Systems with Applications, 231: #120646 [DOI: 10.1016/j.eswa.2023.120646]

Li Y Z, Chang M C and Lyu S W. 2018. In Ictu Oculi: exposing AI created fake videos by detecting eye blinking//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630787]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020b. Celeb-DF: a large-scale challenging dataset for DeepFake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Liu H G, Li X D, Zhou W B, Chen Y F, He Y, Xue H, et al. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 772-781 [DOI: 10.1109/CVPR46437.2021.00083]

Luo Y C, Zhang Y, Yan J C and Liu W. 2021. Generalizing face forgery detection with high-frequency features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16312-16321 [DOI: 10.1109/CVPR46437.2021.01605]

Masi I, Killekar A, Mascarenhas R M, Gurudatt S P and AbdAlmageed W. 2020. Two-branch recurrent network for isolating deepfakes in videos//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 667-684 [DOI: 10.1007/978-3-030-58571-6_39]

Pang G L, Zhang B P, Teng Z, Qi Z G and Fan J P. 2023. MRE-Net: multi-rate excitation network for deepfake video detection. IEEE Transactions on Circuits and Systems for Video Technology, 33(8): 3663-3676 [DOI: 10.1109/TCSVT.2023.3239607]

Pu W B, Hu J, Wang X, Li Y Z, Hu S, Zhu B, et al. 2022. Learning a deep dual-level network for robust DeepFake detection. Pattern Recognition, 130: #108832 [DOI: 10.1016/j. patcog. 2022. 108832]

Qian Y Y, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in frequency: face forgery detection by mining frequency-aware clues//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]

Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++ : learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Sabir E, Cheng J X, Jaiswal A, AbdAlmageed W, Masi I and Natarajan P. 2019. Recurrent convolutional strategies for face manipulation detection in videos//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Long Beach, USA: IEEE: 80-87

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2): 336-359 [DOI: 10.1007/s11263-019-01228-7]

Shang Z H, Xie H T, Zha Z J, Yu L Y, Li Y and Zhang Y D. 2021. PRRNet: pixel-region relation network for face forgery detection. Pattern Recognition, 116: #107950 [DOI: 10.1016/j.patcog.2021.107950]

Thies J, Zollhöfer M and Nießner M. 2019. Deferred neural rendering: image synthesis using neural textures. ACM Transactions on Graphics, 38(4): #66 [DOI: 10.1145/3306346.3323035]

Tzaban R, Mokady R, Gal R, Bermano A and Cohen-Or D. 2022. Stitch it in time: GAN-based facial editing of real videos//Proceedings of 2022 SIGGRAPH Asia 2022 Conference Papers. Daegu, Korea (South): ACM: #29 [DOI: 10.1145/3550469.3555382]

Wang G J, Jiang Q, Jin X, Li W and Cui X H. 2022a. MC-LCR: multimodal contrastive classification by locally correlated representations for effective face forgery detection. Knowledge-Based Systems, 250: #109114 [DOI: 10.1016/j.knosys.2022.109114]

Wang H Y, Liu Z H and Wang S L. 2023. Exploiting complementary dynamic incoherence for DeepFake video detection. IEEE Transactions on Circuits and Systems for Video Technology, 33(8): 4027-4040 [DOI: 10.1109/TCSVT.2023.3238517]

Wang J K, Wu Z X, Ouyang W H, Han X T, Chen J J, et al. 2022b. M2TR: multi-modal multi-scale transformers for Deepfake detection//Proceedings of 2022 International Conference on Multimedia Retrieval. Newark, USA: ACM: 615-623 [DOI: 10.1145/3512527.3531415]

Xia Z M, Qiao T, Xu M, Zheng N and Xie S C. 2022. Towards deepfake video forensics based on facial textural disparities in multi-color channels. Information Sciences, 607: 654-669 [DOI: 10.1016/j.ins.2022.06.003]

Xu Y B, Yin Y Q, Jiang L M, Wu Q Y, Zheng C Y, Loy C C, et al. 2022. TransEditor: transformer-based dual-space gan for highly controllable facial editing//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7673-7682 [DOI: 10.1109/CVPR52688.2022.00753]

Yang W Y, Zhou X Y, Chen Z K, Guo B F, Ba Z J, Xia Z H, et al. 2023. AVoiD-DF: audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18: 2015-2029 [DOI: 10.1109/TIFS.2023.3262148]

Yu P P, Fei J W, Xia Z H, Zhou Z L and Weng J. 2022. Improving generalization by commonality learning in face forgery detection. IEEE Transactions on Information Forensics and Security, 17: 547-558 [DOI: 10.1109/TIFS.2022.3146781]

Yu Y, Zhao X H, Ni R R, Yang S Y, Zhao Y and Kot A C. 2023. Augmented multi-scale spatiotemporal inconsistency magnifier for generalized DeepFake detection. IEEE Transactions on Multimedia, 25: 8487-8498 [DOI: 10.1109/TMM.2023.3237322]

Zhang D C, Li C Y, Lin F Z, Zeng D and Ge S M. 2021. Detecting deepfake videos with temporal dropout 3DCNN//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI: 1288-1294 [DOI: 10.24963/ijcai.2021/178]

Zhao C R, Wang C T, Hu G S, Chen H N, Liu C and Tang J H. 2023. ISTVT: interpretable spatial-temporal video transformer for Deepfake detection. IEEE Transactions on Information Forensics and Security, 18: 1335-1348 [DOI: 10.1109/TIFS.2023.3239223]

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021a. Multi-attentional deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2185-2194 [DOI: 10.1109/CVPR46437.2021.00222]

Zhao T C, Xu X, Xu M Z, Ding H, Xiong Y J and Xia W. 2021b. Learning self-consistency for deepfake detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15003-15013 [DOI: 10.1109/ICCV48922.2021.01475]

Zhao X H, Yu Y, Ni R R and Zhao Y. 2022. Exploring complementarity of global and local spatiotemporal information for fake face video detection//Proceedings of 2022 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 2884-2888 [DOI: 10.1109/ICASSP43922.2022.9746061]

Zheng Y L, Bao J M, Chen D, Zeng M and Wen F. 2021. Exploring temporal coherence for more general video face forgery detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15024-15034 [DOI: 10.1109/ICCV48922.2021.01477]

Zhu C T, Zhang B L, Yin Q L, Yin C X and Lu W. 2024. Deepfake detection via inter-frame inconsistency recomposition and enhancement. Pattern Recognition, 147: #110077 [DOI: 10.1016/j.patcog.2023.110077]

Zhu K M, Xu W B, Lu W and Zhao X F. 2022. Deepfake video detection with feature interaction amongst key frames. Journal of Image and Graphics, 27(1): 188-202 (祝恺蔓, 徐文博, 卢伟, 赵险峰. 2022. 多关键帧特征交互的人脸篡改视频检测. 中国图象图形学报, 27(1): 188-202) [DOI: 10.11834/jig.210408]

作者简介

龙敏，女，教授，主要研究方向为数字水印和数字取证。
E-mail：caslongm@aliyun.com
彭飞，通信作者，男，教授，主要研究方向为数字水印和数字取证。E-mail：eepengf@gmail.com
尹茜，女，硕士研究生，主要研究方向为多媒体安全和多媒体取证。E-mail：22108021612@stu.csust.edu.cn
张乐冰，男，副教授，主要研究方向为数字取证、多媒体安全和深度学习。E-mail：zhanglebing@hhtc.edu.cn