基于预训练模型跨层注意力注入机制的语音伪造检测方法

刘斯鸿, 雷震春, 钱广源, 周 勇, 刘长红

江西师范大学人工智能学院 南昌 中国 330022

摘要 近年来, 深度学习的快速进步推动了预训练模型的广泛应用, 尤其在标注数据稀缺的情况下, 其重要性越来越凸显。相关研究表明, 源自这些模型的特征在语音深度伪造检测任务中具有出色的判别能力, 为下游分类任务提供了更加稳健和丰富的信息表示。然而, 大多数现有方法主要利用预训练模型最顶层特征, 从而忽略了较低层和中间层中存在的潜在有价值的信息。尽管近期部分研究尝试在不同层之间传播或融合特征, 但这些方法通常缺乏一种动态机制来促进顶层与底层之间的有效信息流动, 限制了对预训练模型表征能力的充分挖掘。在本文中, 我们提出了一种新颖的跨层注意力注入方法来进一步增强来自预训练模型特征的判别能力。具体而言, 我们并不是简单地使用来自最顶层特征或采用递归融合策略, 而是引入了一种基于注意力的机制。我们首先将最顶层特征选择性地传递到若干较低层, 从而获得一系列语义增强特征表达。为了避免全层无差别传递带来的冗余, 我们对顶层特征进行均值池化来获得全局语义表示, 并通过注意力机制将其注入到这些语义增强特征中以实现有效指导。最终来自最顶层的特征和得到全局语义表示的特征进行门控融合后被输入到精心设计的后端分类器中。在后端, 我们提出了具有多尺度并行建模能力的 Des2Net_BiMamba 模型, 该模型采用了 Res2Net 的并行多分支结构, 其中由多个 Des2Net 层和双向 Mamba 块组成, 它们分别用于多尺度特征提取和建模全局上下文信息。该架构旨在同时捕捉区分真实语音与伪造语音所需的细粒度和整体模式。我们在广泛使用的 ASVspoof 2019 逻辑访问、ASVspoof 2021 逻辑访问、ASVspoof 2021 深度伪造和 IN-THE-WILD 数据集上进行了全面实验, 以评估我们提出的方法的有效性。实验结果表明, 我们的方法在 ASVspoof 2019 逻辑访问、ASVspoof 2021 逻辑访问、ASVspoof 2021 深度伪造和 IN-THE-WILD 数据集上保持高度竞争力。这项研究结果验证了跨层注意力注入方法在语音伪造检测任务中充分发挥预训练模型潜力方面的有效性。消融实验和对比实验分别证明了我们的后端分类器的有效性和跨层注意力注入方法的有效性与便捷性。另外可视化分析进一步验证了我们方法在提升特征判别性方面的优势。总体而言, 我们的研究结果凸显了跨层注意力注入在充分发挥预训练模型潜力方面的巨大潜力,所提出方法显著提升了语音伪造检测的鲁棒性和准确性, 在多个主流的数据集上取得了优异表现。

关键词 语音伪造检测; 跨层注意力注入; 预训练模型; Mamba; Transformer

中图法分类号 TP391; TP183 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.11.13

Cross-Layer Attention Injection from Pre-Trained Models for Spoof Speech Detection

LIU Sihong, Lei Zhenchun, QIAN Guangyuan, ZHOU Yong, LIU Changhong

School of Artificial Intelligence, Jiangxi Normal University, Nanchang 330022, China

Abstract In recent years, the rapid progress of deep learning has driven the widespread adoption of pre-trained models, whose significance has become increasingly prominent, especially in scenarios with limited labeled data. Studies have shown that features derived from these models exhibit remarkable discriminative power in speech deepfake detection tasks, providing more robust and informative representations for downstream classification. However, most existing methods mainly utilize features from the topmost layer of pre-trained models, thereby neglecting potentially valuable information present in lower and intermediate layers. Although some recent works have explored propagating or fusing features across different layers, these approaches typically lack a dynamic mechanism to promote effective information flow between the top and bottom layers, thus limiting the full potential of pre-trained model representations. To address these limitations, we propose a novel Cross-Layer Attention Injection (CLAI) method to further enhance the discriminative capability of features derived from pre-trained models. Specifically, instead of simply using the topmost features or adopting a recursive fusion strategy, we introduce an attention-based mechanism. We first selectively transmit the topmost features to several lower layers, obtaining a series of semantically enhanced feature representations. To avoid the redundancy caused by individual scriminate transmission to all layers, we apply mean pooling to the topmost features to obtain a global semantic representation, which is then injected into these enhanced features via an attention mechanism for effective guidance. The

通信作者: 雷震春, 博士, 副教授, Email: zhenchun.lei@hotmail.com。

本课题得到国家自然科学基金项目(No. 62567003)资助。

收稿日期: 2025-06-30; 修改日期: 2025-10-10; 定稿日期: 2025-10-30

final features, obtained by gated fusion of the topmost features and the global semantic representation, are then fed into the carefully designed backend classifier. In our backend classifier, we propose the Des2Net_BiMamba model, which incorporates parallel multi-scale modeling through a multi-branch architecture inspired by Res2Net. The model consists of multiple Des2Net layers and bidirectional Mamba blocks, which are respectively used for multi-scale feature extraction and modeling global contextual information. This architecture is designed to capture both fine-grained and holistic patterns necessary for distinguishing between bona fide and spoofed speech.We conduct comprehensive experiments on the widely used ASVspoof 2019 Logical Access, ASVspoof 2021 Logical Access, Deepfake and IN-THE-WILD datasets to evaluate the effectiveness of our approach. Experimental results demonstrate that our method remains highly competitive on the ASVspoof 2019 Logical Access, ASVspoof 2021 Logical Access, Deepfake and IN-THE-WILD dataset. Ablation and comparative studies validate the contributions of the backend classifier and the CLAI mechanism, respectively. In addition, visualization analyses confirm the superiority of our method in enhancing the discriminability of feature representations. Overall, our findings underscore the great potential of cross-layer attention injection in fully leveraging pre-trained models, significantly improving the robustness and accuracy of speech deepfake detection, and achieving outstanding performance on multiple benchmark datasets.

Key words spoof speech detection; cross-layer attention injection; pre-trained models; Mamba; Transformer

1 引言

随着深度神经网络的不断发展, 文本到语音(Text-To-Speech, TTS)[1] 和语音转换(Voice Conversion, VC)[2] 技术变得越来越强大。由此, 某些不法分子能够利用合成语音进行非法活动, 对自动说话人验证(Automatic Speaker Verification, ASV)[3-4]系统构成了严重威胁。为了保障语音应用的安全, 语音伪造检测(Spoof Speech Detection, SSD)[5-6]——主要关注识别由 TTS 和 VC 模型生成的合成语音, 已成为 ASV 的重要研究课题。

SSD 的实现流程通常包括两个主要部分: 前端特征提取器和后端分类器。传统前端特征提取方法中, 手工设计的声学特征和原始波形的直接表示被广泛使用。例如 Tomilov 等人[7]采用了短时傅里叶变换(Short-Time Fourier Transform, STFT)对语音信号进行时频分析, 提取出相应的特征用于语音伪造检测。该方法能够在时间和频率上对信号进行局部分析, 因此在许多反欺骗系统中得到广泛使用。Sahidullah 等人[8]提出了一种基于线性频率滤波器组的线性频率倒谱系数(Linear Frequency Cepstral Coefficients, LFCC), 该方法采用线性间隔的滤波器组替代了传统的 Mel 尺度滤波器组。与常用的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)相比, LFCC 在高频部分表现出更强的特征提取能力, 因此在相关语音处理任务中得到了广泛关注。再后来 Lei 等人[9]基于 LFCC 特征在高斯混合模型(Gaussian Mixture Model, GMM)各个分量上的得分分布信息来构建对数高斯概率(Log-Gaussian Probability, LGP)特征, 这种特征强调了语音特征帧之间的局部关系。在此基础上, Cao 等人[10]对原有的 LGP 进行了扩展, 提出了多阶高斯概率特征融合方法, 以促进不同阶数 LGP 特征之间的信息交互, 进一步提升了特征的表达能力。此外, 另一类研究则从分析粒度入手, 例如 Yang 等人[11]提出了包括恒定 Q 等子带变换(Constant-Q Equal Subband Transform, CQ-EST)与倍频程子带变换(Constant-Q Octave Subband Transform, CQ-OST)在内的多种子带特征, 证明通过子带分解比传统的全频带分析能更有效地捕获伪造线索。除了在特征工程上进行创新, 研究者也利用 LFCC 等传统特征来解决更为复杂的泛化性问题。例如, Xie 等人[12]在 LFCC特征之上, 通过设计一种聚合分离域泛化的训练策略, 来提升模型对未见攻击类型的检测能力, 展示了传统声学特征在先进学习框架下的新潜力。

但近年来, 端到端语音处理技术的发展推动了研究者逐步摆脱对传统手工特征的依赖, 使其能够直接在原始音频上优化模型, 并展现出巨大的应用潜力。因此, 目前大量研究工作聚焦于这类端到端模型。例如, Tak 等人[13]首次在语音伪造检测中应用 RawNet2 深度神经网络, 它可以通过时域卷积直接对原始音频进行处理, 实现了有效的特征学习。随后自监督学习(Self-supervised Learning, SSL)作为一种在有限标注数据条件下进行语音表征学习的有力方法逐渐兴起。基于 SSL 的预训练模型(Pre-trained Models, PTMs), 如 Wav2vec2.0, WavLM 和 HuBERT, 在包括 SSD 在内的多种语音处理任务中表现出色。这些模型作为有效的特征提取器, 为构建鲁棒的语音伪造检测系统奠定了坚实基础, 并在 ASVspoof 数据集上取得了可喜的成果。例如, 在 ASVspoof 2021 的逻辑访问(Logical Access, LA)场景和深度伪造(Deepfake, DF)检测任务中, Tak 等人[14]采用了改进后的 Wav2vec2.0 模型对原始波形进行特征提取, 并将这些通用特征输入后端分类器进行识别, 显著提升了分类性能。Pan 等人[15]对 WavLM 模型的结构进行了深入分析, 通过融合来自多个 Transformer 编码器的嵌入特征, 并将其输入到分类器中用于语音伪

造检测任务, 实验结果显著提升。

在后端分类器中, AASIST, Conformer 和 Res2Net 等先进模型因其强大的表征能力和在语音伪造检测任务中的有效性而备受关注。例如, AASIST 利用图注意力网络来处理复杂的局部和全局特征[16], Conformer 结合了卷积和 Transformer 模块[17], 能够更好地捕捉局部与全局信息, Res2Net 则是在 ResNet 的基础上, 通过将特征通道分组并在每个残差块内分层处理, 增强了多尺度特征表征[18]。尽管这些先进方法在语音伪造检测任务中表现优异, 但它们的特征输入通常来自预训练模型的最顶层 Transformer 表征, 这可能会忽略低层和中间层蕴含的宝贵多层次信息。为了解决这一局限, 递归特征学习(Recursive Feature Learning, RFL)方法[19]将预训练模型最顶层特征递归地反馈到低层, 旨在细化各层表征, 随后通过加权融合的方式将顶层特征和这些递归层的特征进行融合, 为下游分类任务提供更具上下文感知的特征。但同时也带来了一些问题, 首先, 递归回传过程中, 信息在逐层传递时容易逐步衰减和混叠, 导致特征信号减弱甚至丢失; 其次, 这种方式往往需要依次经过每一层(如 1、2、3 等), 缺乏跨层灵活性和选择性, 可能带来特征冗余, 从而影响融合特征的判别能力和有效性。

为克服上述不足, 我们提出了一种新颖的跨层注意力注入(Cross-Layer Attention Injection, CLAI)机制。与递归回传不同, 我们首先将最顶层特征选择性地传递到若干较低层, 从而获得一系列语义增强特征表达。为了避免全层无差别传递带来的冗余, 我们对顶层特征进行均值池化来获得全局语义表示, 并通过注意力机制将全局语义表示注入到这些语义增强特征中以实现有效指导, 从而强化与全局语义相关的特征表达, 同时抑制无关信息的干扰, 进一步提升特征的表征能力和判别性能。最终顶层特征与得到全局语义表示的特征进行门控融合后被输入到后端分类器中。在后端部分, 我们提出了具有多尺度并行建模能力的 Des2Net_BiMamba 模型, 该模型采用了 Res2Net 的并行多分支结构, 其中主要由多个 Des2Net 层和多个双向 Mamba(BiMamba)块组成, 分别用于提取多尺度和建模全局上下文信息。综上, 前端与后端的协同设计在特征表达与时序建模方面表现出显著的互补性, 进一步促进了系统整体判别性能的提升。最后在多个主流数据集上的实验结果表明, 我们的方法具有显著的有效性和良好的泛化能力, 取得了优越的性能

2 相关工作

2.1 预训练模型 Wav2vec2.0

Wav2vec2.0[14]是一种先进的自监督语音特征预训练模型, 由 Facebook AI Research 提出。与传统手工特征方法不同, Wav2vec2.0 能够直接在原始音频波形上进行建模, 大幅提升了语音特征的表征能力。其模型架构主要包括卷积特征编码器(CNN Feature Encoder)和 Transformer 编码器, 前者用于提取局部特征, 后者则对序列进行全局建模。在训练过程中, Wav2vec2.0 采用对比学习策略, 通过随机遮蔽部分音频片段, 迫使模型学习到更具判别性的特征。

具体而言, 卷积特征编码器由多层一维卷积组成, 对原始音频序列 x 进行如下变换, 得到潜在特征表示 z:

$$z = f_{\text{conv}}(x) = \{z_1, z_2, \ldots, z_T\} \tag{1}$$

其中, T 表示卷积和下采样后的时间步数, $z_t \in \mathbb{R}^d$ $(1 \le t \le T)$是第 t 帧的特征向量, 其表示如下:

$$z_t = \sigma\left(W^{(l)} * \sigma\left(\ldots\sigma\left(W^{(1)} * x + b^{(1)}\right)\ldots\right) + b^{(l)}\right) \tag{2}$$

其中, $W^{(l)}$ 和 $b^{(l)}$ 分别为第 l 层的卷积核和偏置, *表示卷积操作, σ 为激活函数。

在自监督预训练阶段, 模型对特征序列 z 中的部分时间步进行随机掩码, 得到掩码后特征 $\tilde{z}$:

$$\tilde{z} = \begin{cases} [\text{MASK}], & t \in \mathcal{M} \\ z_t, & t \notin \mathcal{M} \end{cases} \tag{3}$$

其中, $\mathcal{M}$ 为被掩码的时间步集合。随后, 掩码后的特征序列 z 被输入到 Transformer 编码器中, 利用多头自注意力机制对全局上下文进行建模。每层的自注意力机制计算如下:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

$$Q = \tilde{z}W^Q, \quad K = \tilde{z}W^K, \quad V = \tilde{z}W^V \tag{5}$$

$$c = \text{Transformer}(\tilde{z}) \tag{6}$$

其中, d_k 表示每个注意力头中键向量的维度, $c = \{c_1, c_2, \ldots, c_T\}$为 Transformer 的输出特征序列, 每一帧 c_t $(1 \le t \le T)$都融合了全局上下文信息。

在对比学习阶段, 模型最大化上下文表示 c_t 与其对应正样本 q_t 的相似度, 同时最小化与负样本 q_k 的相似度。其对比损失函数定义为

$$\mathcal{L}\text{contrast} = -\sum_{t\in\mathcal{M}} \log \frac{\exp(\text{sim}(c_t, q_t)/\kappa)}{\exp(\text{sim}(c_t, q_t)/\kappa) + \sum k = 1^K \exp(\text{sim}(c_t, q_k)/\kappa)} \tag{7}$$

其中, sim(,)为点积, κ 为温度参数。正样本 q_t 通常指与当前表示 c_t 属于同一类别或来源的样本(如同一说话人、同一音频片段的不同增强版本等), 而负样本 q_k 则指与 c_t 不同类别或来源的其他样本, 用于提升特征表示的判别能力。

2.2 Mamba

Mamba[20]是一种新兴的序列建模架构,它属于“线性状态空间模型(State Space Model, SSM)[21]”的高效变体, 广泛应用于自然语言处理、音频建模、时间序列预测等多种任务。与 Transformer 等基于自注意力机制的模型相比[22-24], Mamba 以更低的计算复杂度实现了对长距离依赖的高效建模, 近年来在大规模预训练与下游任务中取得了优异的性能表现。Mamba 的核心思想是利用参数化的状态空间方程对序列进行递推建模, 其基本结构如下:

$$\begin{aligned} h_{t+1} &= Ah_t + Bx_t \\ y_t &= Ch_t + Dx_t \end{aligned} \tag{8}$$

其中, x_t 表示输入序列在时刻 t 的特征向量, h_t 为状态向量, y_t 为输出, A,B,C,D 为可学习参数。相比传统 SSM, Mamba 通过结构设计和高效实现, 使得状态转移与输出计算可在 GPU 上并行, 极大提升了计算效率。

具体到 Mamba 的实现, 其对状态转移矩阵 A 进行了简化和门控参数化, 递推公式为

$$h_{t+1} = \lambda_t \odot h_t + (1-\lambda_t) \odot (K_t \odot x_t) \tag{9}$$

其中, $\lambda_t = \sigma(W_\lambda x_t + b_\lambda)$ 为门控参数, K_t 为可学习的卷积核, $\odot$ 表示逐元素乘法, σ 为 Sigmoid 激活函数。输出端的映射表达为

$$y_t = W_y h_t + b_y \tag{10}$$

其中, W_y 表示输出层的权重矩阵, 用于将隐藏状态 h_t 映射到输出空间。b_y 表示输出层的偏置项。

Mamba 模型由多层高效的状态空间单元堆叠而成, 其核心的选择性扫描机制配合归一化和非线性激活函数, 能够有效地捕获长距离依赖关系。与 Transformer 的自注意力机制相比, Mamba 因其递归式的计算范式, 计算复杂度呈线性增长, 显著降低了内存消耗和推理延时, 适合大规模数据和实时应用场景。此外, Mamba 在模型的可扩展性和部署灵活性方面也表现突出, 使其能够更好地适应复杂多变的实际需求。近年来, Mamba 在语言建模、语音识别、时间序列预测等众多任务中取得了相当甚至超越 Transformer 的效果, 展现出强大的长序列建模能力与广阔的应用前景。

3 方法概述

如图 1 所示, 所提出的 SSD 系统架构主要包括前端特征提取器和后端分类器两个部分。前端特征提取器以预训练模型 XLS-R 为基础, 首先利用 CNN Feature Encoder 和 Transformer 多层逐步抽取原始音频的多层级特征。为提升特征表达能力, 系统在前端引入了创新的 CLAI 机制。该机制首先对顶层(第 24 层)特征进行均值池化以生成一个全局语义表示(G), 然后我们采用一种近似对数尺度采样策略, 通过注意力模块对不同层级(如第 1、2、4、8、12 层)特征注入全局语义表示。随后得到全局语义表示的特征与顶层自身的原始高级特征一同被送入

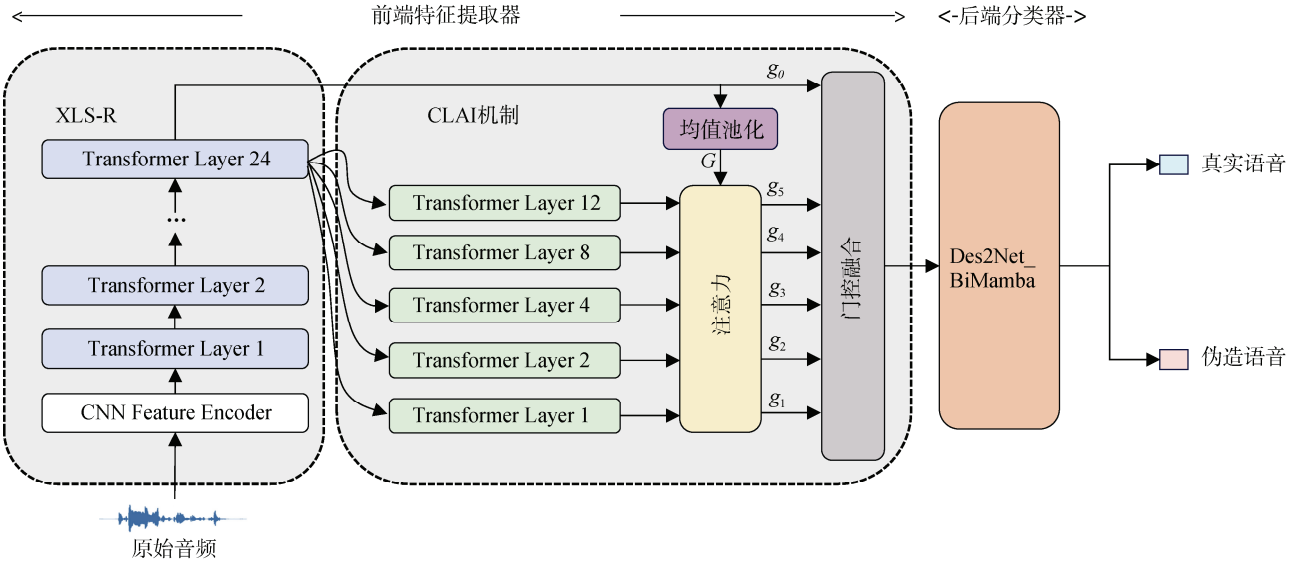

图 1 总体模型图

Figure 1 The structure of the proposed SSD model

后续的门控融合模块, 进行最终的多层次信息融合。随后在后端, 本文提出了具有多尺度并行建模的 Des2Net_BiMamba 模型(详见 3.2 节), 该模型将多尺度局部特征与全局序列特征相结合, 以进一步提升对语音伪造检测任务中复杂特征的捕获和判别能力。最终, 分类器对输入语音进行判别, 并在输出端分别给出真实语音与伪造语音的检测结果。

我们 CLAI 机制中采用的近似对数尺度采样策略, 其设计哲学根植于自监督预训练模型所普遍遵循的特征层次性原理[25-26], 即低层(如 1、2、4 层)通常编码更多声学底层信息(如与频谱相关的声学线索), 中层(如 8、12 层)则包含更丰富的语言学特征, 尤其在音素区分、词身份以及一定的语义信息上表现突出, 高层(如顶层)则进一步编码更抽象、更全局以及与具体任务相关的语义信息。我们的策略正是为了以最小的冗余高效捕获到从底层到中层的完整特征谱系。与之相比, 常规的采样策略则存在固有缺陷。例如连续采样(如 1、2、3 等)会因引入过多高度相关的相邻层特征而导致严重的信息冗余, 并且在获取中高层特征时会带来巨大的计算开销; 而等差采样(如 1、4、7 等)则会因其机械的固定步长, 忽略了特征演化的非线性规律, 从而无法实现高效的最优层组合。

3.1 预训练模型中的跨层注意力注入

XLS-R[27]是一种基于 Wav2vec2.0 框架的大规模自监督语音表征模型, 专为跨语言和多语种场景设计。该模型不仅继承了 Wav2vec2.0 在自监督语音特征学习方面的优越性能, 还通过引入更为丰富和多样化的语言资源, 以及显著扩展的训练数据规模(覆盖 128 种语言, 累计 43.6 万小时的语音数据), 极大地提升了模型的表达能力和泛化能力。XLS-R 的训练语料涵盖了全球主流及低资源语言, 使其能够捕捉到不同语言间的共性与差异性特征, 从而学习到更具普适性和鲁棒性的语音表征。

这种大规模跨语言自监督训练策略, 使 XLS-R 在面对不同语种、口音和说话风格时具备更强的适应能力, 为多语种语音识别、语音合成、语音伪造检测等任务提供了坚实的基础。相关研究表明, XLS-R 在多语种语音任务中的表现优于传统的单语种模型, 尤其在低资源语言场景下具有显著优势。在本研究中, 我们选用 XLS-R 作为 SSD 任务的主干特征提取器, 充分利用其强大的跨语言表征能力, 以提升模型在多样化语音数据上的泛化性能。

我们提出了 CLAI 机制, 旨在充分利用自监督 XLS-R 模型不同层次的特征表示。与常规“取某层输出”不同, 我们首先将最顶层特征作为输入, 分别送入选定的若干 Transformer 层, 获得每个分支的特征。随后, 通过注意力模块和门控融合模块, 实现全局与局部信息的动态融合。对于每个被选中的层, 其分支特征定义为

$$H^{l} = \mathrm{Layer}_{l}(H^{(N)}) \tag{11}$$

其中, $\mathrm{Layer}_{l}(.)$表示第 l 层 Transformer, $H^{(N)}$为 XLS-R 顶层输出。然后对顶层输出 $H^{(N)}$做时序均值池化, 得到全局引导特征:

$$G = \mathrm{MeanPool}(H^{(N)}) \tag{12}$$

随后每个分支特征 H^{l} 通过注意力机制与全局引导特征 G 交互, 增强特征判别力。注意力注入后的特征表示为

$$\widetilde{H}^{l} = H^{l} + \alpha^{l} \cdot \mathrm{Inject}(H^{l}, G) \tag{13}$$

其中, α^{l} 是一个可学习的标量门控, 用于控制全局语义信息注入的程度。Inject(.)表示注意力模块, 实现如下:

$$\mathrm{Inject}(H^{l}, G) = \sum_{t=1}^{T} \mathrm{Softmax}\left(Q \cdot K_{t}^{\top}\right) \cdot V_{t} \tag{14}$$

其中, $Q=W_{q}G$, $K_{t}=W_{k}H_{t}^{l}$, $V_{t}=W_{v}H_{t}^{l}$, W_{q}, W_{k}, W_{v} 为线性变换矩阵。

所有注入后的分支特征 $\widetilde{H}^{(l)}$ 与顶层特征 $H^{(N)}$, 通过可学习门控系数进行自适应融合, 最终特征如下公式所表示:

$$X = \sum_{i \in S} g_{i} \cdot \widetilde{H}^{(l)} + g_{0} \cdot H^{(N)} \tag{15}$$

其中, S 表示包含一组经过选择的较低层特征索引的集合, g_{i} 是每个分支特征的门控参数, g_{0} 是顶层特征的门控参数。

CLAI 机制使最终表征能够通过基于注意力的全局语义表示注入, 从 XLS-R 模型顶层以及若干较低层整合互补信息。通过融合不同深度的特征, CLAI 机制能够有效利用 Transformer 中编码的底层与顶层语义线索, 从而在语音伪造检测任务上获得更具信息量和判别性的语句级表征。这一机制不仅增强了特征表达的丰富性和区分性, 还提升了模型对多样化伪造攻击的鲁棒性和泛化能力, 使其在复杂的实际应用场景中表现更加优异。

3.2 Des2Net_BiMamba 模型

如图 2 所示, Des2Net_BiMamba 模型采用了一种类似 Res2Net 的并行多分支结构, 从而具备了强大的多尺度并行建模能力。具体来说, 我们在每个 Des2Net 层内部引入有效模块以提取多尺度特征(详见 3.2.1 节), 并在每个分支中融合带有残差缩放的 BiMamba 块, 以

实现高效的全局上下文建模(详见 3.2.2 节), 共同增强模型的表征能力。

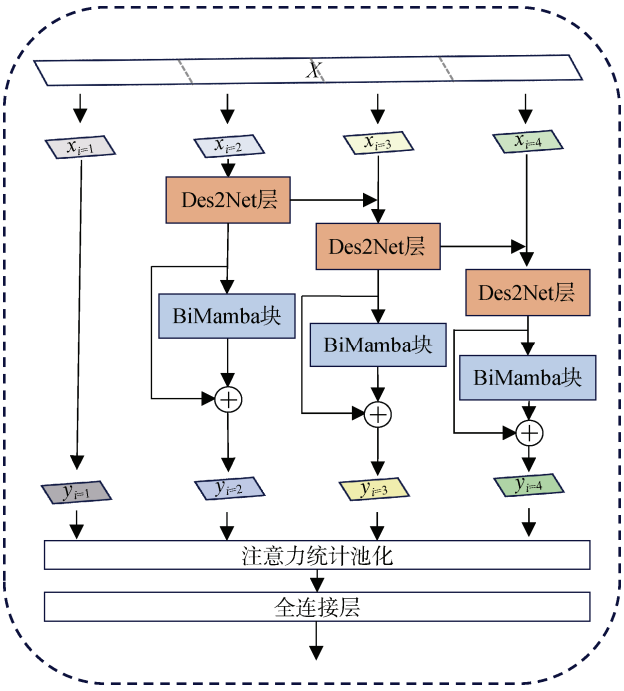

图 2 Des2Net_BiMamba 模型结构

Figure 2 The structure of Des2Net_BiMamba

对于输入特征图 $X \in \mathbb{R}^{C\times T}$, 首先在通道维度上均匀划分为 s 个子集(图中以 $s=4$ 为例), 记为 x_i, 其中 $i \in \{1, 2, \cdots, s\}$。这样的划分方式能够有效提升特征处理的灵活性与表达能力。对于第一个分支($i=1$), 输出等于输入, 即 y_1=x_1, 无须进一步处理, 实现信息的直接传递。对于其余分支($i \geq 2$), 子集 x_i 首先经过 Des2Net 层以提取多尺度特征, 随后通过 BiMamba 块进一步建模全局信息, 并引入可学习的残差缩放。各分支的输出定义如下:

$$h_i = \begin{cases} 0, & i=1 \\ \text{Des2Net}(x_i + h_{i-1}), & 2 \leq i \leq s \end{cases} \tag{16}$$

$$z_i = h_i + \gamma \cdot \text{BiMamba}(\text{LN}(h_i)) \tag{17}$$

$$y_i = \begin{cases} x_i, & i=1 \\ z_i, & 2 \leq i \leq s \end{cases} \tag{18}$$

其中, 当 $i=1$, $h_i=0$ 表示该分支不经过 Des2Net 层, LN(.)表示层归一化, γ 为可学习的缩放参数。在我们的实验中, γ 被初始化为 0.5。最后得到的输出 y_i 统一经过注意力统计池化和全连接层来生成最终的检测结果。

3.2.1 Des2Net 层

如图 3 所示, 每个 Des2Net 层旨在于各分支内提取丰富且多样的特征。输入 x_i, 首先通过一个 1×1 卷积(为简洁起见, 公式中省略), 并在通道维度上进一步划分为 s' 个分组(图中以 $s'=4$ 为例), 记为 $x_{i,j}$, 其中 $j \in \{1, 2, \cdots, s'\}$。对于每个分组, 通过深度可分离卷积(Depthwise Separable Convolution, DSC)提取多尺度特征, 并通过动态加权求和(Dynamic Weight Sum, DWS)机制进行动态融合。第 j 个分组的操作如下:

$$s_{i,j} = \text{DSC}_j(x_{i,j} + y_{i,j-1}), \quad y_{i,0} = 0 \tag{19}$$

$$\hat{s}_{i,j} = \text{DWS}_j(s_{i,j}) = s_{i,j} \cdot \left[\alpha_j \cdot \sigma\left(\text{AvgPool}(s_{i,j})\right)\right] \tag{20}$$

其中, α_j 为可学习参数, $\sigma(.)$表示 Sigmoid 激活函数, AvgPool(.)表示全局平均池化操作, 随后所有分组的特征经过求和后, 投影回原始通道维度:

$$Y_i = \text{Conv}_{1\times 1}\left(\sum_{j=1}^{s'} \hat{s}_{i,j}\right) \tag{21}$$

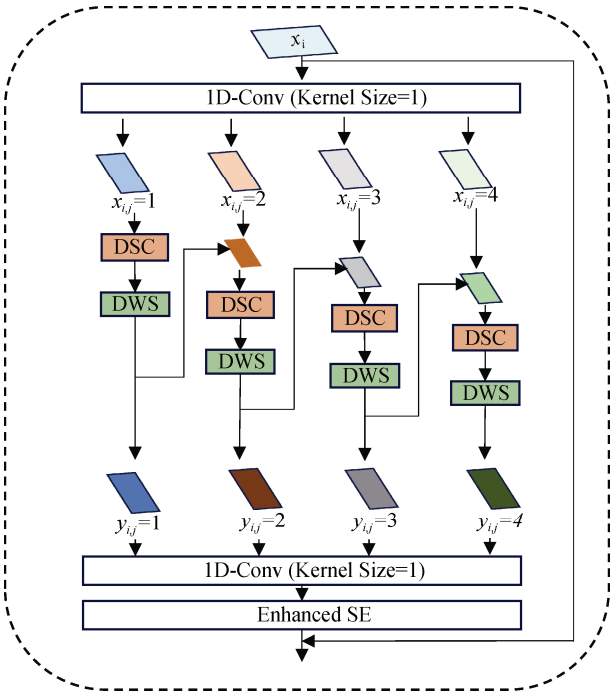

图 3 Des2Net 层

Figure 3 The structure of Des2Net Layer

为了进一步增强特征表征, 在每个 Des2Net 层中集成了增强型挤压-激励模块(Enhanced Squeeze and Excitation, Enhanced SE)。Enhanced SE 通过将全局平均池化和最大池化的结果拼接作为通道描述符对通道特征进行重新校准:

$$c_i = \text{Concat}\left(\text{AvgPool}(Y_i), \text{MaxPool}(Y_i)\right) \tag{22}$$

平均池化能够捕捉输入特征的全局统计信息, 有助于模型理解整体分布和趋势, 提升对全局结构的感知。最大池化则突出每个通道最显著的激活响应, 使模型关注关键局部特征和异常。两者结合不仅

兼顾整体与细节, 还提升了对不同类型特征的表达和判别能力。

最后增强特征 c_i 和原始特征 x_i 通过残差连接相加, 以保留原始信息并提升模型的训练稳定性和泛化能力。

3.2.2 BiMamba 块

BiMamba[28]块是在 Mamba 架构基础上进行改进设计的, 旨在通过双向序列建模更充分地捕捉全局特征依赖关系。如图 4 所示, 具体来说, BiMamba 在原有单向 Mamba 分支的基础上, 新增了一条反向建模分支, 从而实现对输入序列的前向和后向建模。对于输入特征, BiMamba 分别通过两条并行的 Mamba 分支进行处理, 其中前向分支直接对原始序列进行建模, 后向分支则对序列进行反向(Reverse)操作后输入。通过这种双向结构, BiMamba 能够融合正向和逆向的全局上下文信息, 有效提升模型对复杂序列依赖关系的捕捉能力, 为后续特征判别提供更丰富的信息支持。对于输入 $h_i \in \mathbb{R}^{C\times T}$, 分别通过两条并行的 Mamba 分支进行前向和后向序列建模, 其中后向分支的输入为序列进行反向(Reverse)操作后的结果:

$$u_{i,f} = \text{Mamba}(h_i), \quad u_{i,b} = \text{Mamba}(\text{Reverse}(h_i)) \tag{23}$$

两条分支的输出(后向分支输出再 Reverse 回来)通过相加进行融合:

$$u_i = u_{i,f} + \text{Reverse}(u_{i,b}) \tag{24}$$

最终输出 u_i 乘以一个可学习的残差缩放参数, 并通过残差连接和 Des2Net 层的输出进行相加, 最后通过 ASTP 和全连接层进行分类。

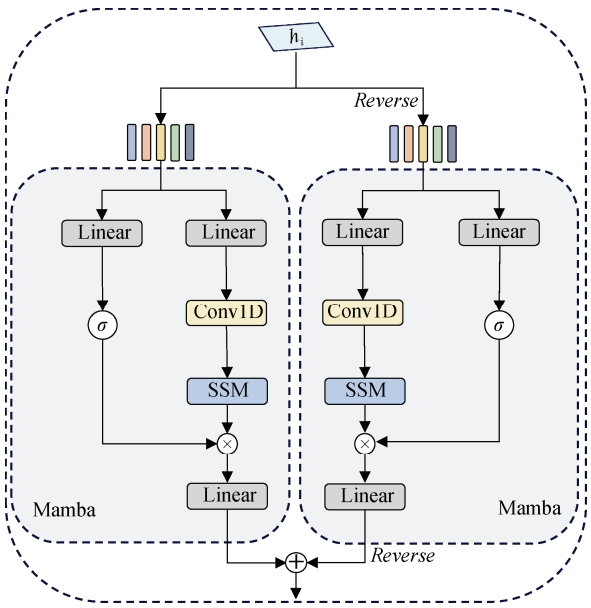

图 4 BiMamba 块

Figure 4 The structure of BiMamba Block

4 实验设置

4.1 数据集

为验证所提方法的有效性, 我们严格遵循 ASVspoof 官方评测协议: 在 ASVspoof 2019 LA 的训练集上训练模型, 并在其开发集上进行超参数选择与验证。训练完成后, 将模型分别在 ASVspoof 2019 LA 评估集、ASVspoof 2021 LA、ASVspoof 2021 DF 和 IN-THE-WILD(ITW)数据集上进行测试。值得注意的是 ASVspoof 2021 LA 与 ASVspoof 2021 DF 数据集仅提供评估集, 不包含公开的训练或开发集。四个数据集的类别与样本分布详见表 1。

表 1 ASVspoof 2019 LA、ASVspoof 2021 LA、ASVspoof 2021 DF 及 ITW 数据集分布

Table 1 The distribution of ASVspoof 2019 LA, ASVspoof 2021 LA, 2021 DF and ITW datasets

	bonafide(真实语音)	spoof(伪造语音)
19 LA 训练集	2580	22800
19 LA 开发集	2548	22296
19 LA 评估集	7355	63882
21 LA	18452	163114
21 DF	22617	589212
ITW	19963	11816

ASVspoof 2019 LA 数据集基于 VCTK 语音库构建, 涵盖了真实语音和多种类型的伪造语音样本。伪造语音主要通过 TTS 和 VC 技术生成。在训练集和开发集中, 伪造语音由编号 A01～A06 的算法产生, 包括 2 种语音转换算法和 4 种语音合成算法。评估集中的伪造语音由编号 A07-A19 的算法生成, 涵盖了 6 种语音转换和 7 种语音合成方法。值得注意的是, 评估集包含两种与训练集相同的算法, 其余均为训练阶段未见过的未知算法, 这为模型的泛化能力评估提供了有力依据。ASVspoof 2021 LA 数据集覆盖了更加丰富的伪造场景, 共包含约 18 万条语音, 主要分为三类: 目标说话人的真实语音、基于 TTS 系统合成的伪造语音以及基于 VC 系统转换的伪造语音。该数据集囊括了多种 TTS 和 VC 系统, 涉及不同的声码器和建模方法, 包括基于深度神经网络的先进合成与转换系统, 大幅提升了伪造语音的自然度和多样性。ASVspoof 2021 DF 数据集包含约 60 万条语音数据, 涵盖 100 种伪造及真实语音类型。该数据集中的伪造语音通常不依赖传统的文本或说话人信息, 直接采用端到端的深度生成模型合成, 具有高度自然、难以分辨的特性, 在音质和说话人特征上与真实语音极为相似, 从而对自动说话人验证系统

提出了更高的挑战。此外, 所有音频样本均经过多种无损编解码器处理, 进一步引入了压缩多样性, 为语音防伪系统的评测提供了更加具有挑战性和现实意义的基准。ITW 数据集包含 20.7 小时的真实数据和 17.2 小时的伪造数据。与传统 ASVspoof 系列数据集不同, 该数据集直接采集自真实网络环境, 其中真实语音采集自同一演讲者的播客和演讲, 伪造音频来源于 219 个公开分享的深度伪造宣传演讲视频。这使得该数据集更加贴近实际应用中的开放环境, 具有更强的多样性和挑战性。

4.2 评估指标

性能评估方面我们采用 ASVspoof 2021 挑战赛主办方提供的最小串联检测代价函数(Minimum tandem Detection cost Function, min t-DCF)[29]和等错误率(Equal Error Rate, EER)[29]作为主要指标。min t-DCF 和 EER 的值越低代表模型的性能越好。

4.2.1 最小串联检测代价函数

为评估语音伪造检测模型与 ASV 系统联合应用时的整体性能, ASVspoof 2019 引入了以 ASV 为中心的串联检测代价函数(Tandem Detection Cost Function, t-DCF)。该指标能够同时反映声纹识别系统与反伪造系统的综合性能表现, 为了便于计算和横向比较, 通常采用该指标的最小归一化形式, 其定义如下式所示:

$$\min t-\mathrm{DCF}=\min_{\tau_{cm}}\left\{\frac{C_0+C_1P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})+C_2P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})}{C_0+\min\{C_1,C_2\}}\right\} \tag{25}$$

其中, $P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})$ 和 $P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})$ 分别表示 CM 系统在阈值 τ_{cm} 处的未命中率和误报率; C_0, C_1, C_2 都取决于 t DCF 参数和 ASV 系统的错误率。它们的表示如下所示, 具体参数设置通常依据实际应用场景进行调整, 以确保评估的公平性和科学性。

$$\begin{aligned}C_0&=\pi_{\mathrm{tar}}C_{\mathrm{miss}}P_{\mathrm{miss}}^{\mathrm{asv}}+\pi_{\mathrm{non}}C_{\mathrm{fa}}P_{\mathrm{fa}}^{\mathrm{asv}}\\C_1&=\pi_{\mathrm{tar}}C_{\mathrm{miss}}-\left(\pi_{\mathrm{tar}}C_{\mathrm{miss}}P_{\mathrm{miss}}^{\mathrm{asv}}+\pi_{\mathrm{non}}C_{\mathrm{fa}}P_{\mathrm{fa}}^{\mathrm{asv}}\right)\\C_2&=\pi_{\mathrm{spoof}}C_{\mathrm{fa,spoof}}P_{\mathrm{fa,spoof}}^{\mathrm{asv}}\end{aligned} \tag{26}$$

其中, π_{tar}, π_{non}, π_{spoof} 分别表示目标说话人、非目标说话人和伪造语音的先验概率(非负且总和为 1)。C_{miss}, C_{fa}, $C_{\mathrm{fa,spoof}}$ 分别表示漏检目标用户、错误接收非目标用户说话人和错误接受伪造语音的代价。$P_{\mathrm{miss}}^{\mathrm{asv}}$, $P_{\mathrm{fa}}^{\mathrm{asv}}$, $P_{\mathrm{fa,spoof}}^{\mathrm{asv}}$ 表示在 ASV 系统阈值下未命中率、误报率和伪造攻击错误接受率。

4.2.2 等错误率

等错误率(Equal Error Rate, EER)是一项常用于评估语音伪造检测系统性能的重要指标。EER 指的是当系统的误拒率(False Rejection Rate, FRR)与误接率(False Acceptance Rate, FAR)相等时对应的错误率数值。FRR 反映了真实语音被误判为伪造语音的概率, 而 FAR 则表示伪造语音被误判为真实语音的概率。EER 越低, 通常代表检测模型在真实与伪造语音的区分能力越强, 整体性能越好。

4.3 实现细节

实验基于 PyTorch 框架, 并使用 RTX 4090 GPU 进行。训练阶段中, 音频数据被裁剪或拼接为约 4 秒(64,600 采样点)的片段。我们采用 AdamW 优化器, 初始学习率设为 1×10^{-5}, 权重衰减系数为 1×10^{-4}, 并优化加权交叉熵损失函数。所有实验的批量大小均为 20。为提升收敛速度和模型稳定性, 我们采用学习率调度器(ReduceLROnPlateau): 当验证损失在连续 4 个迭代周期内未下降时, 学习率按 0.1 的因子降低, 最低学习率设为 1×10^{-7}。特征图 $X\in\mathbb{R}^{C\times T}$ 的通道数 C 设为 1024。当验证集交叉熵损失在 8 个迭代周期内无提升时, 实验停止。最终性能通过对验证集表现最优的前 5 个模型进行模型平均后得到的模型检查点进行报告。

此外, 在训练过程中, 我们采用了先进的 RawBoost[30]数据增强技术, 以模拟多种信道和编解码器效应, 从而显著增加训练数据的多样性与真实性。RawBoost 是一种专为语音反欺骗和鲁棒性研究设计的波形级数据增强方法。其核心思想是通过在原始音频波形上施加多种信号扰动, 通过添加不同的干扰噪声对训练数据进行不同程度的增强。其信号扰动类型丰富, 包括线性和非线性卷积噪声、与脉冲信号相关的加性噪声和与稳态信号无关的加性噪声, 以及它们之间的组合, 能够覆盖更广泛的真实场景变化。这一增强策略不仅有效提升了模型的泛化能力, 还使模型能够更好地适应各种复杂和多变的实际应用环境, 进一步增强了模型对复杂声学条件下输入信号的处理能力, 最终提高了对不同类型伪造攻击的鲁棒性。

5 实验结果及分析

5.1 在多个数据集上的性能对比

如表 2 所示, 我们给出了各系统在 ASVspoof 2021 LA 数据集上针对不同攻击类型(A07–A19)的 EER 结果。前三个基线系统均采用了 Wav2vec2.0 模型最顶层的输出作为后端分类器的输入特征, 而 W2V2+RFL[19]则引入了递归特征学习方法以进一步提升预训练模型的特征表达能力。对比结果显示, 无论传统基线方法还是引入 RFL 的改进系统, 在部分

表 2 ASVspoof 2021 LA 数据集上的性能表现, 其中 A07 至 A19 代表用于 LA 评估的 13 种攻击算法

Table 2 The performance on the ASVspoof 2021 LA dataset, where A07 to A19 refer to 13 attack algorithms used for LA evaluation

模型	ASVspoof 2021 LA													pooled EER(%)	min t-DCF
	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19		
W2V2+Liner layer	0.41	0.45	0.28	8.79	9.06	0.98	0.22	0.34	1.13	0.80	0.54	1.61	0.58	2.45	0.2549
W2V2+AASIST[14]	0.39	0.56	0.31	1.17	1.07	0.59	0.21	0.32	0.41	0.49	0.48	1.94	0.45	0.87	0.2081
W2V2+MFA[31]	0.28	0.45	0.16	0.63	0.86	0.36	0.18	0.27	0.32	0.40	0.38	1.52	0.51	0.63	0.2019
W2V2+RFL[19]	0.23	0.33	0.10	0.38	0.49	0.27	0.10	0.12	0.24	0.28	0.30	1.32	0.28	0.47	0.1975
Proposed	**0.22**	**0.16**	**0.09**	**0.16**	**0.12**	**0.12**	**0.03**	**0.08**	**0.12**	**0.10**	**0.21**	**0.67**	**0.23**	**0.24**	**0.1916**

攻击类型下仍存在检测性能的提升空间。总体来看, 我们提出的方法在每一种攻击类型下均取得了最优的EER表现, 充分验证了所设计系统的有效性和鲁棒性。特别是在整体 EER(pooled EER)指标上, 相比W2V2+RFL 方法实现了 48.9%的显著降幅, 突出展现了模型在多样化伪造攻击场景下的强大适应能力。

为进一步与最新的 SSD 系统进行比较, 表 3 展示了所提出方法与当前最先进系统在 ASVspoof 2019 LA 评估集、ASVspoof 2021 LA、2021DF 和 ITW 这四个数据集上的性能对比。如表 3 所示, 我们的方法在 ASVspoof 2019 LA 评估集和 ASVspoof 2021 LA 上均取得了最优的性能, EER 分别为 0.05%与 0.24%, min t-DCF 分别为 0.0015 与 0.1916。值得注意的是, 近年来基于自监督预训练模型的方法表现强劲, 这主要得益于其能够利用大规模外部真实语音语料库。尽管如此, 我们提出的方法在 ASVspoof 2021 LA 上得到的 EER 和 min t-DCF 均取得了进一步的提升, 超越了目前主流的 SSD 系统, 包括 W2V2+RFL[19] (EER 0.47%), W2V2+MFA[31] (EER 0.63%)以及 W2V2+Mamba[32] (EER 0.93%)。这些对比结果进一步验证了所提方法的有效性, 为语音伪造检测领域的持续发展提供了有价值的参考。

表 3 在 ASVspoof 2019 LA 评估集、ASVspoof 2021 LA、ASVspoof 2021 DF 和 ITW 数据集上与当前最先进系统的性能对比

Table 3 Performance comparison with state-of-the-art systems on the ASVspoof 2019 LA_eval, ASVspoof 2021 LA, ASVspoof 2021 DF and ITW datasets

年份	模型	ASVspoof 2019 LA 评估集		ASVspoof 2021 LA		ASVspoof 2021 DF	ITW
		EER(%)	min t-DCF	EER(%)	min t-DCF	EER(%)	EER(%)
2022	W2V2+AASIST†[14]	0.21#	0.0063#	1.00	0.2120	3.69	10.46
2023	W2V2+Conformer[17]	0.24	0.0080	1.38	0.2216	2.58	8.42
2024	W2V2+Conformer+TCM[35]	0.20#	0.0047#	1.03	0.2131	2.25	7.79
2024	W2V2+OCKD[36]	0.39	-	0.91	0.2079	2.07	7.68
2024	W2V2+AASIST2[37]	0.15	-	1.61	-	2.77	-
2024	W2V2+STJ-GAT+BLDL*[38]	0.06	0.0018	0.56	0.2000	1.89	-
2024	W2V2+SLS†[39]	-	-	3.88	-	2.09	7.46
2024	W2V2+MFA[31]	-	-	0.63	0.2019	2.26	-
2025	W2V2+MOE[40]	0.74	-	2.96	-	2.54	9.17
2025	W2V2+Mamba[32]	-	-	0.93	0.2081	1.88	6.71
2025	LSR+LSA[33]	0.15	-	1.19	-	2.43	5.92
2025	LSR+LSA*[33]	0.12		1.05	-	1.86	5.54
2025	W2V2+Nes2Net-X†[34]	-	-	2.00	-	**1.78**	**5.52**
2025	W2V2+RFL[19]	0.08	-	0.47	0.1975	-	-
2025	W2V2+STCA+LMDC[41]	0.09	0.0028	0.78	0.2057	1.87	-
	Proposed	**0.05**	**0.0015**	**0.24**	**0.1916**	2.03	6.83

(注: †表示来自三次独立训练的平均结果, *带有额外的数据增强, #代表我们复现的结果)

此外, 虽然我们的方法在 ASVspoof 2021 DF 和 ITW 数据集上未取得最优的 EER, 但仍实现了较低的错误率和稳定的性能。我们分析认为, DF 条件下的伪造语音样式更加多样和复杂, 不同方法在

特征建模和判别机制上存在差异。例如, LSA[33]和Nes2Net-X[34]等方法对于DF任务具有更强的针对性, 但这也在一定程度上牺牲了它们在 LA 场景下的性能。相比之下, CLAI 机制通过多层级注意力融合, 有效增强了模型对 LA 任务中关键伪造特征的感知能力; Des2Net_BiMamba 结构采用多尺度特征提取与全局上下文建模, 更好地捕捉了 LA 场景下语音信号的多层次变化与伪造痕迹, 这使得所提方法在某种程度上更“专精”于 LA 任务, 而导致在 DF 任务上泛化能力的小幅折中。ITW 数据集不同于 ASVspoof 系列数据集, 其更加突出真实场景的多样性和不可预测性, 因此为模型评估带来了更大挑战, 另外有些伪造样本还可能存在“部分伪造”等更具挑战性的情况。这些因素对模型的泛化能力和鲁棒性提出了更高要求。综合来看, 该方法在以上四个数据集上都取得了优异且均衡的表现, 这说明该方法在不同数据分布和复杂攻击场景下依然保持了稳定的性能优势, 具有较高的实际应用潜力。不过, 需要指出的是, 在更贴合真实场景的 ITW 数据集和复杂多样的 DF 数据集上, 模型表现仍有提升空间。面对更丰富的语音内容和伪造手段, 提升模型的泛化能力和鲁棒性, 依然是后续亟须深入研究和关注的重要研究方向。

5.2 消融实验

表 4 展示了所提各个模块在 ASVspoof 2019 LA 评估集和 ASVspoof 2021 LA 数据集上的消融实验结果。通过对比完整系统与去除不同模块后的性能, 可以更清楚地理解每个模块在整体架构中的实际作用。实验发现, 无论去除哪个模块, 系统性能都会出现不同程度的下降, 这从侧面反映了各模块的不可替代性。

表 4 所提模块的消融实验结果

Table 4 Ablation results of the proposed modules

模型	2019 LA 评估集	2021 LA
	EER(%)	EER(%)
去除 CLAI	0.22	1.20
去除 BiMamba	0.12	0.43
去除 Enhanced SE	0.16	0.55
去除 DWS	0.08	0.32
Proposed	**0.05**	**0.24**

具体来看, CLAI 模块对模型性能的提升作用尤为显著。当该模块被移除后, 在 20119LA 评估集上和 2021LA 数据集上, 系统的 EER 分别由 0.05%升至 0.22%, 0.24%升至 1.20%。这说明 CLAI 机制能够有效整合多层信息, 帮助模型捕捉更丰富的特征, 是提升系统判别能力的关键环节。类似地, 去除 BiMamba 时序建模模块后, EER 分别上升至 0.12%和 0.43%。BiMamba 通过对语音序列的时序特征建模, 增强了系统对时变特征的敏感性, 其缺失导致模型对伪造语音的识别能力下降。Enhanced SE 模块同样发挥了重要的作用。当该模块被移除时, EER 分别上升至 0.16%和 0.55%, 表明改进的通道注意力有助于模型聚焦于最具判别力的特征通道, 提高了模型特征表达的有效性。此外, DWS 模块能够根据输入自适应地调整各路特征的重要性, 实现信息的高效融合。当移除 DWS 后, EER 分别升至 0.08%和 0.32%, 也体现了其对最终性能的积极贡献。

总的来说, 消融实验结果清楚地揭示了各模块在系统中的独特价值。每个模块的协同工作共同保证了系统的最优性能(EER 分别为 0.05%和 0.24%)。这些结果不仅验证了整体架构设计的合理性和有效性, 也为后续模型优化和实际应用提供了有益参考和指导意义。

5.3 层级特征及采样策略对系统性能的影响

为系统性地研究在注意力机制中引入不同层级的特征对系统性能的影响, 我们设计了一种近似对数尺度采样策略。该策略遵循特征层次性原理, 旨在高效捕获多样化且信息冗余度低的特征谱系。基于该思路, 我们并评估了七种不同的采样策略: ①仅使用顶层特征; ②顶层与第 1 层特征; ③顶层与第 1、2 层特征; ④顶层与第 1、2、4 层特征; ⑤顶层与第 1、2、4、8 层特征; ⑥顶层与第 1、2、4、8、12 层特征; 以及⑦顶层与第 1、2、4、8、12、16 层特征。上述策略涵盖了从低层到高层不同层级特征作为注意力模块输入的组合方式, 旨在全面考察多层级信息对模型表现的影响。

如图 5 所示, 随着更多较低层特征被输入到注意力机制, 系统在语音伪造检测任务上的性能稳步提升, 说明多层级信息为注意力机制提供了更丰富的判别依据, 有助于提升模型的判别能力。尤其值得关注的是, 当输入顶层与第 1、2、4、8、12 层特征时, 系统获得了最低的 EER 和 min t-DCF, 表明在此配置下模型能够充分整合不同层级的关键信息, 实现最优的伪造检测性能。然而, 进一步输入第 16 层特征(策略⑦)时, EER 和 min t-DCF 升高, 系统性能出现下降, 可能是由于引入了冗余或噪声信息, 影响模型的判别效果。该趋势在图 5 的最后一个数据点中表现得尤为明显。

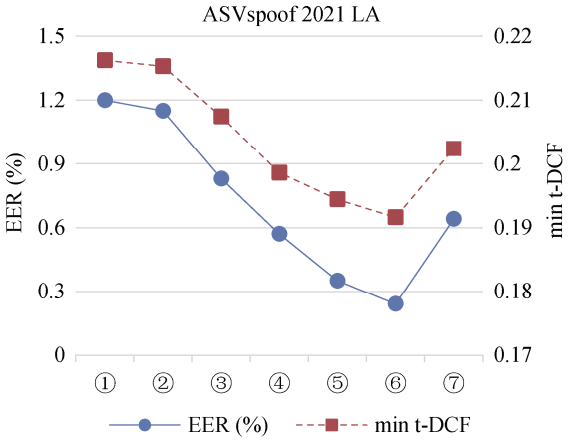

图 5 ASVspoof 2021 LA 数据集上，不同层级特征输入注意力机制时的 EER(%)与 min t-DCF

Figure 5 EER (%) and min t-DCF with different layer feature inputs to the attention mechanism on the ASVspoof 2021 LA dataset.

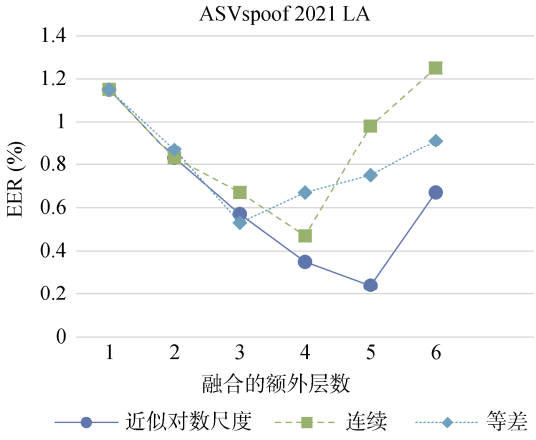

图 6 ASVspoof 2021 LA 数据集上，不同采样策略融合层数下的 EER(%)对比

Figure 6 EER(%) comparison of different sampling strategies with varying number of fused layers on the ASVspoof 2021 LA Dataset.

在此基础上，我们设计了另一个对比实验，以验证近似对数尺度采样策略在固定计算代价下的优越性。实验比较了三种策略(近似对数尺度、连续、等差)在融合相同数量的额外层(1～6 层)时的系统性能。具体配置为：近似对数尺度采样逐步融合{1}, {1、2}, {1、2、4}, {1、2、4、8}, {1、2、4、8、12}, {1、2、4、8、12、16}层；连续采样为{1}, {1、2}, {1、2、3}, {1、2、3、4}, {1、2、3、4、5}, {1、2、3、4、5、6}层；等差采样为{1}, {1、4}, {1、4、7}, {1、4、7、10}, {1、4、7、10、13}, {1、4、7、10、13、16}层，所有实验均在统一设置下进行，保证了结果的可比性。实验结果如图 6 所示，X 轴为融合的额外层数，Y 轴为 EER(%)。可以看到，在大多数配置下，近似对数尺度采样策略都取得了最佳或次优性能，尤其在融合 4 或 5 个层时优势显著。例如，融合 5 个层时 EER 仅为 0.24%，而连续和等差策略分别为 0.98%和 0.75%。

这一结果强有力地证明了我们的采样策略并非偶然有效，而是因为它更好地遵循了预训练模型中特征层次性原理。连续采样因在低层引入过多相似特征而导致冗余；等差采样则因其机械的固定步长，忽略了特征演化的非线性规律，导致其可能在低层错过某些关键细节，同时在中高层会引入过多的冗余信息。相比之下，我们的近似对数尺度采样策略通过在低层密集、高层稀疏的采样方式，实现了对特征多样性与信息冗余的最佳平衡，能够最大限度发挥模型优势，提升系统对复杂语音伪造攻击的鲁棒性和泛化能力。

5.4 不同融合策略性能与计算开销综合评估

为全面评估所提出跨层融合方法的有效性与可行性，我们不仅对现有融合策略进行了详细对比，还系统考察了各融合策略与基线方法在计算开销(参数量、FLOPs、推理延时)方面的表现。所有方法的特征维度均统一设置为 1024，以保证参数规模和计算复杂度的一致性。直接相加、特征拼接以及我们提出的方法均采用了近似对数尺度的采样策略，以充分覆盖不同语义层级的信息融合；递归融合则从第 1 层开始依次堆叠。此外，表 5 第一行所示“顶层特征+简单分类器”作为基线方法，其分类器由池化层和全连接层构成，反映了典型的简化后端设计；后三种融合方法(即直接相加、特征拼接、递归融合)所使用的后端均为 Des2Net_BiMamba，以确保比较的公平性和一致性。需要注意的是，直接相加方法是对选定各层特征进行逐元素相加，不改变输出特征的维度，因而和仅使用顶层特征类似，不会增加前端的参数量；特征拼接则将多层特征在最后一维进行拼接，导致融合后特征维度增加，从而使前端降维线性层的参数量有所增加；递归融合和 CLAI 机制的参数量增加则主要来源于每个分支独立引入的全连接层或注意力模块，这些结构为每条分支都分配了专门的参数，从而进一步提升了模型的特征建模能力。

实验结果如表 5 所示，可以看出，本文所提方法虽然在参数量(333.2 M)、FLOPs(67.15 GMac)和推理延时(11.94 ms)上相较基线方法均有一定幅度的增加，但在 ASVspoof 2021 LA 和 ITW 两个数据集上，所提方法分别取得了 0.24%和 6.83%的 EER，较基线方法

表 5　不同融合策略与基线方法的计算开销及检测性能对比

Table 5　Comparison of Computational Cost and Detection Performance Across Different Fusion Strategies and Baselines

方法	参数量(M)	FLOPs(GMac)	推理延时(ms)	ASVspoof 2021 LA	ITW
				EER(%)	EER(%)
顶层特征+简单分类器	319.0	56.44	9.55	2.68	15.72
顶层特征 +Des2Net_BiMamba	323.8	58.44	10.09	1.20	8.43
Proposed	333.2(333.2)	67.15(67.15)	11.94(11.94)	0.24(0.24)	6.83(6.83)
直接相加	323.8(323.8)	63.44(64.96)	10.69(11.25)	1.06(1.74)	8.02(10.69)
特征拼接	331.1(330.1)	64.53(66.27)	10.82(11.52)	0.93(1.62)	7.91(9.89)
递归融合	329.0(330.1)	64.41(66.49)	11.28(11.61)	0.39(0.68)	7.68(9.26)

(注: 后四行括号外和括号内的结果分别代表在最优融合层数与相同额外层数条件下的性能表现)

分别提升了 91.0%和 56.6%, 这充分展现了本文所提方法在多样化和复杂攻击场景下的强大判别能力和良好泛化性。此外, 相比于直接相加、特征拼接和递归融合这些现有的融合策略, CLAI 机制在检测性能上展现出了显著优势。无论对比相同的额外层数(即表 5 中括号内的结果), 还是对比各自的最优融合层数(即表 5 中括号外的结果), 本文所提方法在两个数据集上始终取得了最低的 EER, 表现出更优且均衡的检测能力。这不仅体现了 CLAI 机制对不同数据分布和复杂攻击场景的强大适应性, 也验证了其在模型泛化性方面的突出表现。这一优势主要得益于 CLAI 机制通过引入注意力模块, 对不同层级的特征注入全局语义信息, 增强了多层特征之间的交互与关键信息的提取; 同时结合近似对数尺度的采样策略, 高效覆盖了从低层到高层的多样化特征谱系, 最大限度减少了信息冗余并充分利用了各层特征的互补性, 从而实现了更加灵活且高效的特征融合与权重自适应, 有效提升了跨层信息的协同建模能力。

整体来看, 尽管所提方法在参数量和计算开销上略有增加, 但整体资源消耗依然可控。在保证参数量和计算开销处于可接受范围的前提下, 所提方法仍能够持续带来性能的提升, 进一步证明了所提方案在实际应用中的高效性和实用价值。

5.5　CLAI 机制在不同后端的表现

鉴于我们框架的模块化设计能够灵活地将特征提取器与后端分类器进行组合, 我们对 CLAI 机制在多种具有代表性的后端分类器上的有效性进行了全面评估, 其中包括 Conformer, AASIST 以及我们所提出的后端分类器 Des2Net_BiMamba。如图 7 所示, 我们在 ASVspoof 2021 LA 上进行实验, 在所有测试的后端中集成 CLAI 机制后均能显著降低 EER, 凸显了该方法的广泛适用性和稳健性。值得注意的是, 在集成 CLAI 机制前, 我们的后端虽具备一定的检测能力, 但 EER 相对较高; 而在集成 CLAI 机制后, 其 EER 降至最低, 展现出最优的检测性能。这一结果不仅说明 Des2Net_BiMamba 与 CLAI 机制的集成效果极为显著, 而且表明所提出的 CLAI 机制能够进一步挖掘模型在语音伪造检测任务中的潜力。

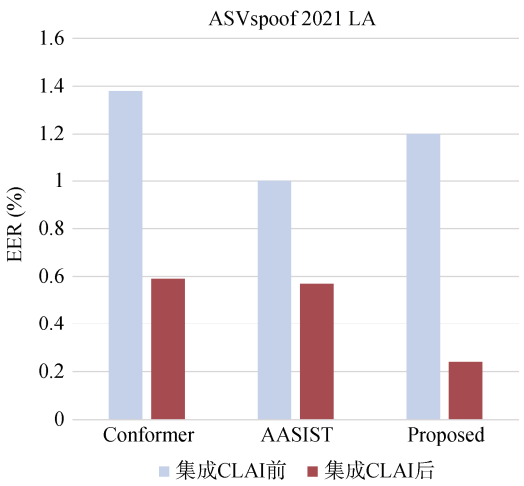

图 7　不同后端模型集成 CLAI 机制前后的性能对比

Figure 7　Performance comparison of different back-end models before and after integrating the CLAI

此外, CLAI 机制在 Conformer 和 AASIST 等不同架构的后端分类器上同样实现了性能提升, 进一步验证了该策略的通用性和适应性。无论后端模型的结构如何变化, CLAI 机制均能有效提升特征的可分性和判别能力, 显著降低伪造语音的误判率。这些发现充分证明, CLAI 机制不仅能够灵活适配多种后端架构, 还能最大限度地提升我们系统的整体性能, 对于构建高鲁棒性、强泛化能力的语音伪造检测系统具有重要意义。

5.6　不同模型潜在特征分布的可视化分析

为更加直观地验证所提方法的有效性和可靠性, 我们在 ASVspoof 2021 LA 数据集上, 采用 t 分布随

机邻域嵌入方法(t-SNE), 对 W2V2+AASIST, W2V2+Conformer, W2V2+RFL 以及本文提出的方法(Proposed)这四个模型最后一层提取的潜在表示进行了可视化, 并对各模型特征分布进行了详细比较。如图 8 所示。值得说明的是, 为保证对比的公平性和可视化结果的代表性, 我们对真实语音(bonafide)与伪造语音(spoof)样本进行了等量采样, 采样比例为 0.2, 即从每一类中分别随机采集 20%的样本用于 t-SNE 分析。

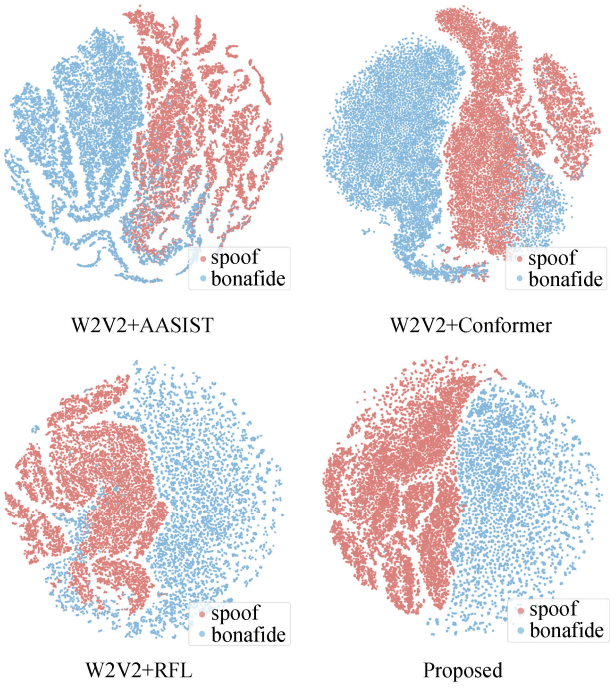

图 8 不同模型在 ASVspoof 2021 LA 数据集上提取的特征表示的 t-SNE 可视化

Figure 8 t-SNE visualization of feature representations extracted by different models on the ASVspoof 2021 LA dataset.

通过对 t-SNE 可视化结果的分析, 可以更直观地观察各模型在区分真实语音与伪造语音样本时的特征分布。可以看出, 三种主流模型在一定程度上能够分离真实与伪造样本, 但嵌入空间中两类样本的分布仍存在明显重叠, 部分类别边界较为模糊, 表明这些模型在特征区分性方面仍有提升空间。而采用所提出的方法后, 嵌入空间中真实与伪造语音的样本分布更加清晰, 并且几乎没有重叠区域。

这种变化不仅说明所提方法能够有效提升模型的判别能力和抗欺骗能力, 也为后续语音伪造检测任务打下了坚实基础。此外, 不同模型间的对比进一步验证了所提方法在多种架构下的适用性和普适性, 表现出良好的泛化能力和应用价值。t-SNE 的可视化结果直观展现了特征空间的优化效果, 充分体现了本方法的优势与实际意义。

6 结论

在本文中, 我们提出了一种新颖的语音伪造检测系统, 面向 ASVspoof 2019 LA、ASVspoof 2021 LA、 2021 DF 和 ITW 数据集。该系统集成了多项先进模块, 包括 CLAI 机制, 通过注意力机制将全局语义表示注入到不同层级特征中, 从而实现对关键信息的增强与无关信息的抑制, 进一步提升了特征的判别性和泛化能力; 此外, 系统后端融合了具有多尺度并行建模的 Des2Net_BiMamba 模型, 其中包含 Des2Net 层以实现多尺度特征提取, 以及 BiMamba 块以增强全局上下文建模能力。系统性的实验表明, 我们的方法在 ASVspoof 2019 LA 评估集、ASVspoof 2021 LA、2021 DF 和 ITW 数据集上表现出较强的竞争力, 充分验证了所提架构的有效性和通用性。消融实验和对比实验分别证明了各个模块的重要性、采样策略的有效性以及 CLAI 机制的便捷性。进一步的可视化分析显示, 所提出方法能够更清晰地区分真实语音与伪造语音, 有效提升了检测任务的准确性和鲁棒性。未来的工作将聚焦于提升该框架在 ITW 等贴近真实场景数据集上的泛化能力, 并进一步增强其在 DF 等复杂多样数据集上的鲁棒性, 以更好地适应实际应用需求。

参考文献

[1] Perrotin O, Stephenson B, Gerber S, et al. Refining the Evaluation of Speech Synthesis: A Summary of the Blizzard Challenge 2023[J]. *Computer Speech & Language*, 2025, 90: 101747.

[2] Sisman B, Yamagishi J, King S, et al. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29: 132-157.

[3] Snyder D, Garcia-Romero D, Sell G, et al. X-Vectors: Robust DNN Embeddings for Speaker Recognition[C]. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018: 5329-5333.

[4] Chen C, Rong Y F, Ji C Q, et al. Speaker Verification Method Based on Deep Information Divergence Maximization[J]. *Journal on Communications*, 2021, 42(7): 231-237.
(陈晨, 彤娅峰, 季超群, 等. 基于深层信息散度最大化的说话人确认方法[J]. *通信学报*, 2021, 42(7): 231-237.)

[5] Wang X, Yamagishi J, Todisco M, et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech[J]. *Computer Speech & Language*, 2020, 64: 101114.

[6] Tao J H, Fu R B, Yi J Y, et al. Development and Challenge of Speech Forgery and Detection[J]. *Journal of Cyber Security*, 2020, 5(2): 28-38.
(陶建华, 傅睿博, 易江燕, 等. 语音伪造与鉴伪的发展与挑战

[J]. *信息安全学报*, 2020, 5(2): 28-38.)

[7] Tomilov A, Svishchev A, Volkova M, et al. STC Antispoofing Systems for the ASVspoof2021 Challenge[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 61-67.

[8] Sahidullah M, Kinnunen T, Hanilçi C. A Comparison of Features for Synthetic Speech Detection[C]. *Interspeech 2015*, 2015: 2087-2091.

[9] Lei Z C, Yan H, Liu C H, et al. Two-Path GMM-ResNet and GMM-SENet for ASV Spoofing Detection[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 6377-6381.

[10] Cao M M, Lei Z C, Yang Y G, et al. Research on Multi-Order GMM-ResNet Fusion for Speech Deepfake Detection[J]. *Journal of Cyber Security*, 2025, 10(2): 116-126.
(曹明明, 雷震春, 杨印根, 等. 多阶 GMM-Res Net 融合在语音伪造检测中的研究[J]. *信息安全学报*, 2025, 10(2): 116-126.)

[11] Yang J C, Das R K, Li H Z. Significance of Subband Features for Synthetic Speech Detection[J]. *IEEE Transactions on Information Forensics and Security*, 2020, 15: 2160-2170.

[12] Xie Y K, Cheng H N, Wang Y T, et al. Domain Generalization *via* Aggregation and Separation for Audio Deepfake Detection[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 344-358.

[13] Tak H, Patino J, Todisco M, et al. End-to-End Anti-Spoofing with RawNet2[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 6369-6373.

[14] Tak H, Todisco M, Wang X, et al. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation[C]. *The Speaker and Language Recognition Workshop* (*Odyssey 2022*), 2022: 112-119.

[15] Pan Z H, Liu T C, Sailor H B, et al. Attentive Merging of Hidden Embeddings from Pre-Trained Speech Model for Anti-Spoofing Detection[C]. *Interspeech 2024*, 2024: 2090-2094.

[16] Jung J W, Heo H S, Tak H, et al. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 6367-6371.

[17] Rosello E, Gomez-Alanis A, Gomez A M, et al. A Conformer-Based Classifier for Variable-Length Utterance Processing in Anti-Spoofing[C]. *INTERSPEECH 2023*, 2023: 5281-5285.

[18] Gao S H, Cheng M M, Zhao K, et al. Res2Net: A New Multi-Scale Backbone Architecture[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(2): 652-662.

[19] Guan Y, Ai Y, Li Z L, et al. Recursive Feature Learning from Pre-Trained Models for Spoofing Speech Detection[C]. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2025: 1-5.

[20] Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces[EB/OL]. 2023: arXiv: 2312.00752. https://arxiv.org/abs/2312.00752.

[21] Gu A, Goel K, Ré C. Efficiently Modeling Long Sequences with Structured State Spaces[EB/OL]. 2021: arXiv: 2111.00396. https://arxiv.org/abs/2111.00396.

[22] Dong L H, Xu S, Xu B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 5884-5888.

[23] Li J H, Duan Z K, Li S R, et al. ESAformer: Enhanced Self-Attention for Automatic Speech Recognition[J]. *IEEE Signal Processing Letters*, 2024, 31: 471-475.

[24] Subakan C, Ravanelli M, Cornell S, et al. Exploring Self-Attention Mechanisms for Speech Separation[J]. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, 2023, 31: 2169-2180.

[25] Pasad A, Chou J C, Livescu K. Layer-Wise Analysis of a Self-Supervised Speech Representation Model[C]. *2021 IEEE Automatic Speech Recognition and Understanding Workshop*, 2022: 914-921.

[26] Vaidya A R, Jain S, Huth A G. Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech[EB/OL]. 2022: arXiv: 2205.14252. https://arxiv.org/abs/2205.14252.

[27] Babu A R, Wang C H, Tjandra A, et al. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale[C]. *Interspeech 2022*, 2022: 2278-2282.

[28] Zhang X Y, Zhang Q Q, Liu H X, et al. Mamba in Speech: Towards an Alternative to Self-Attention[EB/OL]. 2024: arXiv: 2405.12609. https://arxiv.org/abs/2405.12609.

[29] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 47-54.

[30] Tak H, Kamble M, Patino J, et al. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 6382-6386.

[31] Wu H C, Zhang J, Zhang Z T, et al. Robust Spoof Speech Detection Based on Multi-Scale Feature Aggregation and Dynamic Convolution[C]. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2024: 10156-10160.

[32] Xiao Y, Das R K. XLSR-Mamba: A Dual-Column Bidirectional State Space Model for Spoofing Attack Detection[J]. *IEEE Signal Processing Letters*, 2025, 32: 1276-1280.

[33] Huang W, Gu Y M, Wang Z M, et al. Generalizable Audio Deepfake Detection *via* Latent Space Refinement and Augmentation[C]. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2025: 1-5.

[34] Liu T C, Truong D T, Das R K, et al. Nes2Net: A Lightweight Nested Architecture for Foundation Model Driven Speech Anti-Spoofing[EB/OL]. 2025: arXiv: 2504.05657. https://arxiv.org/abs/2504.05657.

[35] Truong D T, Tao R J, Nguyen T, et al. Temporal-Channel Modeling in Multi-Head Self-Attention for Synthetic Speech Detection[C]. *Interspeech 2024*, 2024: 537-541.

[36] Lu J Z, Zhang Y X, Wang W C, et al. One-Class Knowledge Distillation for Spoofing Speech Detection[C]. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics*, *Speech and Signal*

Processing, 2024: 11251-11255.

[37] Zhang Y X, Lu J Z, Shang Z Q, et al. Improving Short Utterance Anti-Spoofing with Aasist2[C]. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2024: 11636-11640.

[38] Wu H C, Guo W, Zhang Z T, et al. Spoofing Speech Detection by Modeling Local Spectro-Temporal and Long-Term Dependency[C]. *Interspeech 2024*, 2024: 507-511.

[39] Zhang Q S, Wen S B, Hu T. Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier[C]. *The 32nd ACM International Conference on Multimedia*, 2024: 6765-6773.

[40] Wang Z Y, Fu R B, Wen Z Q, et al. Mixture of Experts Fusion for Fake Audio Detection Using Frozen Wav2vec 2.0[C]. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2025: 1-5.

[41] Hao Y Q, Xu M Q, Chen Y H, et al. Integrating Spectro-Temporal Cross Aggregation and Multi-Scale Dynamic Learning for Audio Deepfake Detection[C]. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2025: 1-5.

刘斯鸿 于 2023 年在江西师范大学软件工程专业获得工学学士学位。现在江西师范大学人工智能学院软件工程专业攻读硕士学位, 研究领域为语音伪造检测。研究兴趣包括说话人识别、语音情感识别。Email: 202340100338@jxnu.edu.cn

钱广源 于 2023 年在江西农业大学南昌商学院物联网工程专业获得工学学士学位。现在江西师范大学人工智能学院软件工程专业攻读硕士学位, 研究领域为语音情感识别。研究兴趣包括说话人识别、语音伪造检测。Email: qianguangyuan@jxnu.edu.cn

刘长红 于 2011 年在北京科技大学计算机应用技术专业获得博士学位。现任江西师范大学副教授, 硕士生导师。CCF 会员。研究领域为语音-视觉生成、AI+教育、计算机视觉。E-mail: liuch@jxnu.edu.cn

周勇 江西师范大学副教授，硕士生导师。CCF 会员。研究领域为智能信息处理、机器学习。Email: zhouyong@jxnu.edu.cn

雷震春 于 2006 年在浙江大学计算机科学与技术专业获得博士学位。现任江西师范大学副教授, 硕士生导师。CCF 会员。研究领域为说话人识别、语音信号处理。Email: zhenchun.lei@hotmail.com