AWaveFormer: Audio Wavelet Transformer Network for Generalized Audio Deepfake Detection

Rui Wang, Zirui Chen, Bo Wang, *Member, IEEE*, Zhongjie Ba, *Member, IEEE*, and Kui Ren, *Fellow, IEEE*

***Abstract*—Rapid advancements in speech synthesis technology have made it easier to produce realistic synthetic speech, which poses serious threats to public privacy and security. Recent studies have investigated pre-trained models for feature extraction and adopted advanced architectures, including convolutional and graph neural networks, for deepfake detection. Although these methods improve detection performance to some extent, their generalization and robustness still face challenges. To address this issue, in this paper, we propose a forgery detection method based on the fusion of dual pre-trained features and an optimized Transformer architecture. Specifically, we use a cross-attention mechanism to fuse the audio features extracted from pre-trained Wav2vec 2.0 and WavLM, and replace the token mixer in the traditional Transformer architecture with wavelet transform and multi-scale pooling operations. By combining dual pre-trained features, we extract more comprehensive and discriminative features while optimizing the Transformer architecture, which not only reduces time complexity but also enhances detection performance. We conducted experiments on several datasets, and the results show that our model achieves impressive detection performance with EERs of 0.13%, 2.33%, 3.63%, 10.25%, 5.15%, and 0.21% on the ASVspoof 2019LA, ASVspoof 2021LA, ASVspoof 2021DF, In-the-Wild, Fake-or-Real, and ASVspoof 2015LA evaluation datasets, respectively.**

***Index Terms*—Audio deepfake detection, multi-scale learning, Wav2Vec 2.0, WavLM, wavelet transform.**

I. Introduction

With the rapid development of generative technology, synthetically generated modalities, including speech, facial images, textual news, and videos, are widely spread in the real-world and cyberspace [1]. Synthesized content has unexpected implications [2], including distortion of public perception, harm to personal reputation, and escalation of societal unrest or economic losses. Among these threats, deepfake audio has emerged as an urgent issue due to its rapid advancement and unparalleled realism in mimicking human speech [3]. Its technology maturity continues to improve, creating a formidable challenge for existing detection methods [4], [5], [6]. This study focuses on the detection of synthesized speech by exploring advanced, reliable, and computationally efficient methodologies.

To address the potential risks posed by deepfake audio, developing efficient synthesized speech detection systems has become a vital research focus. Early studies predominantly relied on handcrafted audio features, including linear frequency cepstrum coefficients (LFCC) [7], constant Q transform (CQT) [8], and Mel-frequency cepstrum coefficients (MFCC) [9]. These features were typically combined with models such as gaussian mixture models (GMM) [10] and lightweight convolutional neural networks (LCNN) [11]. While these approaches demonstrated satisfactory performance in detecting high-quality forged speech, they also revealed significant limitations. Firstly, handcrafted features often fail to capture comprehensive speech information, inevitably discarding some important information during feature extraction [12], [13], [14]. For example, they typically use fixed-window spectral analysis that limits capturing fast temporal changes, and may overlook subtle spectral details important for detecting synthetic speech. Additionally, these features are largely focused on frequency domain information, offering a narrow perspective that restricts the model's ability. Meanwhile, the traditional detection networks are relatively simple, limiting their capacity to fully leverage the speech information. These limitations underscore the need for more robust and comprehensive approaches that can adapt to the increasing mature of deepfake audio techniques.

In recent years, significant progress has been made in audio forgery detection, with researchers exploring advanced techniques to address the limitations of earlier methods [15], [16], [17], [18]. For instance, fine-tuning pre-trained models has been widely adopted to improve feature expressiveness, while more sophisticated back-end networks, such as AASIST [19] and GMM-ResNet [20], have been proposed to enhance detection performance. Additionally, novel approaches like contrastive learning [21], distillation learning [12], [22], and one-shot learning [23] have been effectively utilized to tackle forgery detection. However, despite these advancements, challenges persist, particularly when it comes to generalizing detection models to handle the complexities and variability of real-world scenarios. These challenges are further compounded by the growing demand for models that balance high detection accuracy with computational efficiency. Consequently, future research directions should focus on improving model robustness and generalization across

Received 18 March 2025; revised 6 July 2025; accepted 11 September 2025. Date of publication 17 September 2025; date of current version 1 October 2025. This work was supported by the National Natural Science Foundation of China under Grant 62372452 and in part by Youth Innovation Promotion Association CAS. The associate editor coordinating the review of this article and approving it for publication was Dr. Zhangjie Fu. *(Corresponding author: Bo Wang.)*

Rui Wang, Zirui Chen, and Bo Wang are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning 116081, China (e-mail: julywang@mail.dlut.edu.cn; chenzirui@mail.dlut.edu.cn; bowang@dlut.edu.cn).

Zhongjie Ba and Kui Ren are with the School of Cyber Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: zhongjieba@zju.edu.cn; kuiren@zju.edu.cn).

Digital Object Identifier 10.1109/TASLPRO.2025.3611229

diverse scenarios, while optimizing for both accuracy and efficiency to meet the practical needs of real-world applications.

In forged face detection, wavelet transform and Transformer architectures have been extensively explored. However, their application in audio deepfake detection remains limited. Synthesized speech can appear choppy and unnatural in time and frequency, whereas natural speech is smoother and more consistent. The wavelet transform can leverage multi-scale analysis to effectively capture unnatural signal characteristics, thereby offering enhanced insight into subtle artifacts indicative of manipulation. Although Transformer shows impressive detection capabilities, its high computational cost poses a challenge for efficiency [24]. Furthermore, most existing audio deepfake detection algorithms rely on a single pre-trained feature, and the limited use of multiple pre-trained features has constrained further improvements in detection performance. To address these challenges, this study proposes a novel approach that integrates wavelet transform with an optimized Transformer architecture and leverages dual pre-trained features to enhance feature representation and improve detection performance.

In this study, we propose a novel audio deepfake detection method. The front-end feature extraction is based on the fusion of dual pre-trained features, while the back-end classification network employs an optimized Transformer architecture. This combination aims to capture both the subtle artifacts in synthesized speech and the natural consistency of genuine speech, leading to a more robust and efficient detection system. Specifically, we select two pre-trained speech models, WavLM [25] and Wav2Vec 2.0 XLS-R (W2V2) [26]. The WavLM model leverages masked speech prediction for automatic speech recognition (ASR) tasks and speech denoising modeling for non-ASR tasks, demonstrating robust cross-task capabilities. W2V2, trained on a large corpus, excels in extracting generalizable and efficient speech features. These self-supervised learning models provide a strong foundation for capturing diverse and representative speech features. To further enhance feature representation, we employ a cross-attention mechanism to fuse the outputs of WavLM and W2V2, extracting more informative features by leveraging their complementary strengths. In the Transformer architecture, we replace the conventional token mixer with a wavelet transform and multi-scale pooling to reduce temporal complexity. The wavelet transform captures multi-scale information across different frequency bands, while multi-scale pooling improves the extraction of features at varying temporal resolutions. These innovations enhance detection performance and computational efficiency. Our approach aims to reduce computational overhead while maintaining high accuracy in detecting synthesized speech, offering a more effective and robust solution to the challenges of audio deepfake detection. The primary contributions of this work are summarized as follows:

- We propose a novel feature fusion framework that leverages a cross-attention mechanism to integrate dual pre-trained features, enabling the extraction of richer and more distinctive representations for improved audio deepfake detection.
- We design an efficient Transformer based classification network enhanced by wavelet transform and multi-scale pooling, which reduces computational overhead while improving artifact detection capabilities across diverse temporal and frequency scales.
- Through extensive experiments on multiple public datasets, we demonstrate that our method achieves state-of-the-art (SOTA) performance, outperforming existing approaches in both detection accuracy and generalization across different scenarios.

This paper is organized as follows. Section II reviews the related works, Section III outlines the proposed detection method, Section IV details the experimental setup, and Section V analyzes the experimental results in detail. Conclusions in Section VI.

II. Related Works

In this section, we take a look at self-supervised pre-trained models for deepfake detection, transformers for deepfake detection, and wavelet transform for deepfake detection tasks.

A. Self-Supervised Pre-Trained Models for Deepfake Detection Tasks

Large-scale pre-trained models have demonstrated powerful feature representation capabilities by leveraging extensive training on massive amounts of unlabeled data. In the domain of audio deepfake detection, self-supervised pre-trained models are increasingly being adopted. Zhang et al. [27] proposed a detection framework that integrates a sensitive layer selection module to extract informative features from multiple layers of the Wav2Vec 2.0 model, thereby enhancing the sensitivity to subtle artifacts in deepfake audio. Xie et al. [28] combined Wav2Vec 2.0 with LCNN and Transformer blocks to effectively capture temporal dependencies and hierarchical contextual cues in audio signals to enhance detection accuracy. Guo et al. [29] employed the pre-trained WavLM model along with a multi-fusion attention classifier, demonstrating promising performance in detecting synthesized speech. Similarly, Li et al. [30] utilized the pre-trained HuBERT model [31], and further enhanced detection performance through a hybrid data augmentation strategy.

In conclusion, the use of self-supervised pre-trained models for deepfake detection has garnered significant attention in audio deepfake detection. However, relying solely on a single pre-trained model often yields limited feature representations, which motivates our study to integrate two distinct pre-trained models, aiming to capture audio features from diverse perspectives and enhance the effectiveness of deepfake detection.

B. Transformer for Deepfake Detection Tasks

The Transformer architecture has gained popularity in classification tasks due to its self-attention mechanism, which efficiently captures long range dependencies in input sequences. It has been successfully applied in audio deepfake detection. For example, Yadav et al. [32] proposed the patched spectrogram synthetic speech detection Transformer, which showcases the effectiveness of patch-wise spectrogram modeling in synthetic speech detection. Cuccovillo et al. [33] introduced a multi-task audio Transformer that performs synthetic speech detection by

analyzing both the magnitude and phase of speech signals. Liu et al. [34] proposed Rawformer, which integrates 2D convolutions with Transformers to effectively capture both local and global dependencies in raw waveform inputs. Shin et al. [35] developed HM-Conformer, a Conformer-based deepfake detection framework that incorporates hierarchical pooling and multi-level classification token aggregation. Li et al. [36] proposed Safear, a novel Transformer-based detection framework that enables effective audio deepfake detection while preserving the privacy of speech content.

Transformer-based architectures have demonstrated strong classification capabilities and have been widely adopted in deepfake detection tasks. However, the high time complexity and insufficient specialization for audio-specific characteristics restrict the applicability of such models. To address these challenges, our work explores a more efficient Transformer design by replacing the traditional token mixer with a wavelet transform and multi-scale pooling strategy. This modification enhances the model's multi-scale feature analysis capability while reducing computational overhead, thereby improving both detection performance and efficiency in audio deepfake scenarios.

C. Wavelet Transform for Deepfake Detection Tasks

The wavelet transform has gained widespread attention and application due to its superior multi-scale analytical capabilities. Inspired by this fact, various studies have investigated the use of wavelet transformations to detect synthesized speech. Xie et al. [37] proposed a wavelet prompt tuning method that adaptively captures type-invariant auditory forgery cues in the frequency domain without introducing additional trainable parameters. This approach achieves promising generalization performance across datasets. Similarly, Pham et al. [38] applied wavelet transform to convert raw input audio into spectrogram representations, upon which several deep learning-based classifiers were evaluated to distinguish real from fake audio samples. Fathan et al. [13] further investigated different types of wavelet decomposition applied to Mel-spectrograms, and proposed the waveletCNN architecture. This model integrates wavelet analysis directly into convolutional layers to enhance spectral feature extraction, demonstrating improved modeling capacity for synthesis-related distortions. Zbezhkhovska et al. [39] incorporated both Sinc and wavelet filters into the feature extraction layers of the RawNet2 architecture, enabling the model to effectively capture subtle artifacts of synthetic audio. Avaiya et al. [40] introduced the scattering transform to extract stable and multi-scale time-frequency representations. This approach increased sensitivity to subtle signal irregularities, thus improving robustness in cross-domain detection tasks.

Previous studies have shown that wavelet transform is effective for detecting audio deepfakes, either by applying it directly to raw audio signals or by integrating it into neural networks. Inspired by these works, we also introduce wavelet transform for deepfake detection. However, unlike methods that apply it to raw audio, we incorporate wavelet transform into the Transformer architecture. This is because our front-end already extracts robust features using pretrained models. Applying wavelet transform at this stage allows the model to further capture multi-scale structures in the high-level features. Meanwhile, integrating wavelet transform as a token mixer within the Transformer allows the model to perform multi-resolution analysis on the pretrained feature space, capturing both global dependencies and fine-grained temporal variations. This design combines the semantic strength of pretrained representations with the multi-resolution analysis ability of wavelets, improving detection accuracy and generalization.

III. Proposed Method

This section introduces the detection framework we proposed, named AWaveFormer, along with its core components and design principles. We represent the original input speech as $\boldsymbol{X} \in \mathbb{R}^{L \times D}$, where L denotes sequence length, and D denotes the number of features.

A. Overview

The general framework, illustrated in Fig. 1, proposes an effective audio deepfake detection method incorporating self-supervised feature extraction, cross-attention fusion, and a multi-scale wavelet Transformer network (MWTN). First, the original audio input $\boldsymbol{X}$ is processed by two self-supervised pre-trained models, WavLM and W2V2. WavLM extracts global semantic features by leveraging its context-aware masking strategy, while W2V2 focuses on capturing local time-frequency patterns through contrastive learning of masked audio segments. Furthermore, the WavLM involves masked speech denoising and prediction, which encourages the model to recover content from corrupted or masked regions, making it highly suitable for spoofed and noisy input scenarios. By combining the detailed local feature representation of W2V2 and the robust global modeling capability of WavLM, the proposed system leverages complementary strengths. Their complementary representations enhance the detection of both global inconsistencies and local artifacts in deepfake audio. These features are then fused through a cross-attention mechanism, which dynamically aligns and combines them, enabling more robust and effective representations. The fused features are input into the MWTN, consisting of positional encoding, multiple multi-scale wavelet pooling modules (MWPMs), and a classification layer. Positional encoding provides temporal information, while MWPM applies wavelet transform and multi-scale pooling to capture features at various time scales and resolutions, enhancing detection accuracy and reducing computational complexity. To optimize efficiency, MWTN replaces the standard multi-head self-attention mechanism with MWPM, reducing time complexity from $O(N^2)$ to $O(N)$. Finally, the processed features are classified through a fully connected network. In summary, this framework integrates comprehensive feature extraction, cross-attention fusion, and multi-scale modeling to enhance both the accuracy and efficiency of audio deepfake detection, providing a robust solution for identifying synthesized speech.

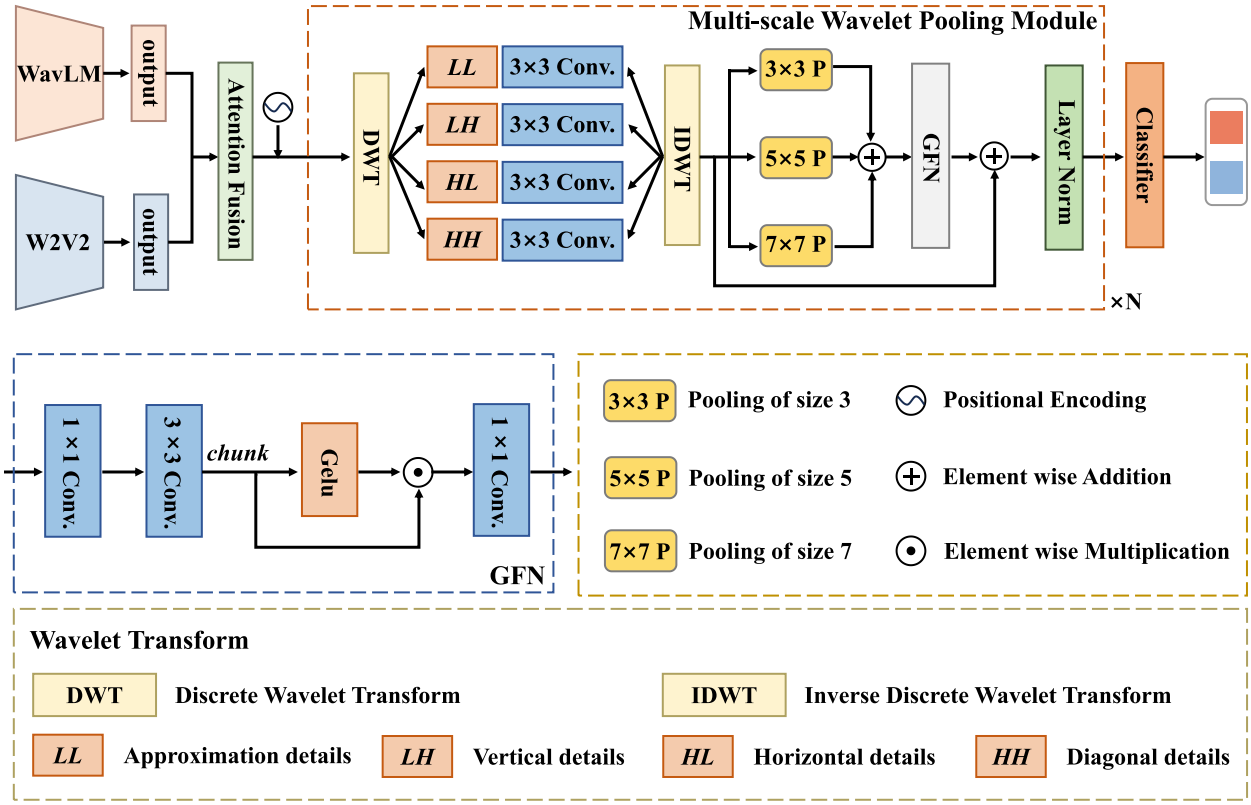

Fig. 1. The structure of the proposed audio wavelet Transformer network for audio deepfake detection. This framework extracts discriminative features from two self-supervised pre-trained models, WavLM and Wav2Vec 2.0, and fuses them using a cross-attention mechanism. These fused features are then processed through positional encoding and passed into N multi-scale wavelet pooling modules for robust synthesized speech detection. The part marked by the blue box in the lower left corner of the picture is the detailed process of the gated feed forward network.

B. Feature Extraction and Fusion

This section introduces the pre-trained self-supervised models used and the cross-attention fusion module.

1) Wav2Vec 2.0: Self-supervised learning has been widely employed in natural language processing and has yielded impressive results. The W2V2 model is a successful implementation of this technology in the speech domain. W2V2 uses self-supervised learning to derive powerful feature representations from raw speech signals, resulting in high performance in specialized applications such as voice recognition and speech forgery detection. Fig. 2 illustrates its pre-training stage, where the original speech signal $\boldsymbol{X}$ is input into the multi-layer convolutional feature encoders for initial feature extraction, producing the feature representation $\boldsymbol{Z}$. The quantization module, combined with the Gumbel Softmax technique [41], discretizes $\boldsymbol{Z}$ into a finite set of phonetic representations q, which are then compared to latent representations in the learning task. Simultaneously, $\boldsymbol{Z}$ is fed into a masked Transformer to capture sequence information, resulting in the final feature representation $\boldsymbol{C}$. By optimizing contrastive loss function, W2V2 learns self-supervised features from unlabeled data, enabling the model to generalize effectively.

In this study, we employed the pre-trained "Wav2Vec 2.0 XLS-R 300 M" model. After loading the official pre-trained parameters, we fine-tuned the model end to end on the audio deepfake detection task. Specifically, we minimized the cross-entropy loss function to optimize the model, enabling it to effectively distinguish between real and synthetic speech. All parameters of the transformer encoder layers were updated during training, allowing the model to adapt its feature representations to better align with the deepfake tasks.

2) WavLM: Unlike W2V2, which relies on high-quality audiobook data, WavLM leverages diverse publicly available audio data, including blogs and YouTube, to better understand real-world speech signals. This diversity enhances WavLM's adaptability, particularly in handling complex, non-standardized speech. Meanwhile, WavLM adds noise and overlaps portions of the input audio, improving both its denoising capabilities and its ability to model speech content. As shown in Fig. 3, during pre-training, WavLM introduces noise or overlap to the input audio signal before feeding it into the convolutional feature encoder for feature extraction. The resulting feature sequence $\boldsymbol{X}$ is then processed by a multi-layer Transformer to capture additional temporal dependencies, producing the final feature representation $\boldsymbol{Z}$. The model is optimized using a mask prediction loss function, which not only enables learning of speech specific features but also captures non-speech attributes related to speaker characteristics and the acoustic environment. This makes WavLM particularly well suited for speech forgery

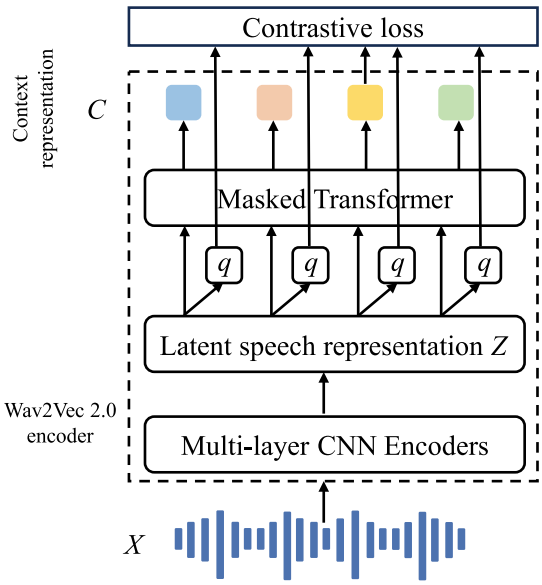

Fig. 2. Illustration of the pre-training process of Wav2Vec 2.0 (Adapted from [26]).

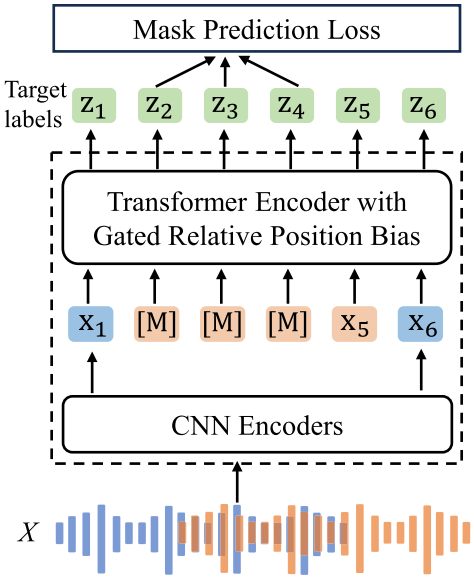

Fig. 3. Pre-training stage of WavLM (Adapted from [25]).

detection, as forged speech often contains artifacts linked to speaker specific traits and unnatural acoustic patterns.

In this work, we employed the pre-trained "WavLM-Large" model. After loading the official pre-trained weights, we fine-tuned all layers of the model on the audio deepfake detection task. This configuration enables the model to extract rich and effective speech features. By leveraging WavLM's powerful self-supervised learning capability and its adaptability to non-standard speech signals, audio artifacts can be effectively detected, ultimately improving the classification performance.

3) Cross-Attention Fusion: Both W2V2 and WavLM produce features at a frame rate of 50 Hz due to their identical convolutional encoder architecture with a total stride of 320.

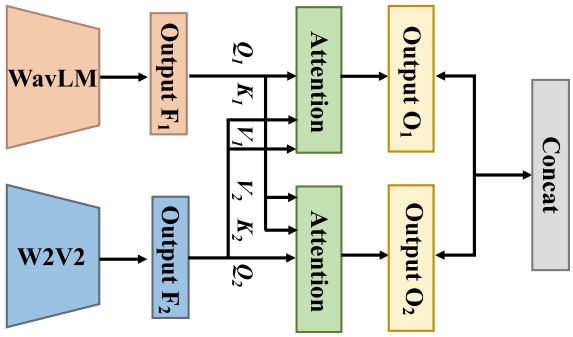

Fig. 4. The structure of the cross-attention fusion. Based on the cross-attention fusion framework, the two self-supervised pre-trained features of WavLM and W2V2 are fused through dot product calculation.

Given the same 16 kHz input, their output sequences are temporally aligned by design, and no additional synchronization is required. Meanwhile, inspired by the effectiveness of the attention mechanism in feature fusion [42], this study introduces a cross-attention method to make more efficient use of pre-trained features. The detailed process is illustrated in Fig. 4. Specifically, the features of the input speech signal are extracted using WavLM, denoted as $\boldsymbol{F}_1$, and W2V2, denoted as $\boldsymbol{F}_2$, where both $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are of dimension 50×64, representing 50 temporal frames and 64-dimensional feature embeddings. In the cross-attention process, $\boldsymbol{F}_1$ is used as the query matrix ($\boldsymbol{Q}_1$), while $\boldsymbol{F}_2$ provides the key ($\boldsymbol{K}_1$) and value ($\boldsymbol{V}_1$). Conversely, $\boldsymbol{F}_2$ serves as the query ($\boldsymbol{Q}_2$), and $\boldsymbol{F}_1$ provides the key ($\boldsymbol{K}_2$) and value ($\boldsymbol{V}_2$). The attention mechanism is computed using the scaled dot-product function. The dot product between the query and key is divided by a temperature coefficient to stabilize the gradient. The similarity score is then normalized using the softmax function to obtain the attention weights. Finally, a weighted sum is performed using the value vectors to generate the fused feature representation. The formula is as follows:

$$\boldsymbol{O}_1 = \text{Softmax}\left(\frac{\boldsymbol{Q}_1 \cdot \boldsymbol{K}_2^\top}{\sqrt{D}}\right)\boldsymbol{V}_2, \tag{1}$$

$$\boldsymbol{O}_2 = \text{Softmax}\left(\frac{\boldsymbol{Q}_2 \cdot \boldsymbol{K}_1^\top}{\sqrt{D}}\right)\boldsymbol{V}_1. \tag{2}$$

The temperature coefficient D is set to 128. The above method produces the feature sequences $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$. Finally, $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ are concatenated along the feature dimension to form the cross-attention fused representation. The cross-attention approach may completely explore the association between W2V2 and WavLM pretrained models, resulting in deep information fusion and classification performance.

C. Multi-Scale Wavelet Transformer Network

The multi-scale wavelet Transformer network (MWTN) is a core component of the model, designed to enhance speech feature processing by combining wavelet transform and multi-scale pooling mechanisms with a gated feed forward network (GFN). Its primary objective is to improve the model's ability to capture

artifacts at various time scales, thereby enhancing feature representation. MWTN consists of three main components, including positional encoding, N multi-scale wavelet pooling modules (MWPMs), and the final classification layer. Each MWPM is composed of three sub-modules, which are wavelet coefficient processing (WCP), multi-scale pooling mechanism (MPM), and GFN. In the following, we provide a detailed description of these modules.

1) Positional Encoding: In the design of the MWTN, we utilize positional encoding to provide the model with sequential information about the input features. This helps the model capture the temporal dependencies in the sequence. Specifically, we use sine and cosine functions for positional encoding, as illustrated below:

$$\boldsymbol{PE}_{pos,2i} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right), \tag{3}$$

$$\boldsymbol{PE}_{pos,2i+1} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right). \tag{4}$$

In these formulas, pos represents the current time step's position, i denotes the index of the positional encoding dimension, and d_{model} refers to the model's feature dimension. The resulting positional encoding is then added to the input features, enabling the model to learn the order relationship within the input sequence. This addition of positional encoding helps the model effectively capture temporal dependencies in audio data, which enhances performance in audio deepfake detection tasks.

2) Wavelet Cofficient Process: In wavelet cofficient process (WCP), the input features are transformed from the time domain to the wavelet domain using the wavelet transform, which helps to enhance the features and capture finer-grained temporal changes. Specifically, the WCP divides the input features $\boldsymbol{F}$, obtained from the cross-attention fusion, into four sub-bands using the discrete wavelet transform (DWT): $\boldsymbol{LL}$ (approximation details), $\boldsymbol{LH}$ (vertical details), $\boldsymbol{HL}$ (horizontal details), and $\boldsymbol{HH}$ (diagonal details). The decomposition can be mathematically expressed as:

$$\boldsymbol{LL}, \boldsymbol{LH}, \boldsymbol{HL}, \boldsymbol{HH} = \text{DWT}(\boldsymbol{F}), \tag{5}$$

$$\begin{aligned}
\boldsymbol{LL}(i,j) &= \sum_{m,n} \boldsymbol{F}(i+m, j+n) \cdot g(m,n), \\
\boldsymbol{LH}(i,j) &= \sum_{m,n} \boldsymbol{F}(i+m, j+n) \cdot h_{\text{LH}}(m,n), \\
\boldsymbol{HL}(i,j) &= \sum_{m,n} \boldsymbol{F}(i+m, j+n) \cdot h_{\text{HL}}(m,n), \\
\boldsymbol{HH}(i,j) &= \sum_{m,n} \boldsymbol{F}(i+m, j+n) \cdot h_{\text{HH}}(m,n),
\end{aligned} \tag{6}$$

$\boldsymbol{F}$ is the input feature, $g[m,n]$ is a low-pass filter, $h_{\text{LH}}[m,n]$, $h_{\text{HL}}[m,n]$, $h_{\text{HH}}[m,n]$ are high-pass filters in the vertical direction, horizontal direction and diagonal direction, respectively, and i, j are the position indices of the input feature. In our implementation, we use the Daubechies 3 (db3) wavelet for both the forward and inverse wavelet transforms. The db3 wavelet offers a good balance between time and frequency localization, and its compact support with three vanishing moments makes it suitable for capturing both transient and stationary features in audio signals.

Then, using convolution, we enhance each coefficient. Convolution helps to reduce the number of parameters while preserving the enhancing effect. For each wavelet coefficient, the convolution process is used for feature extraction as follows.

$$\begin{aligned}
\boldsymbol{Y}_{LL}(i,j) &= \sum_{p,q} \boldsymbol{F}(i+p, j+q) \cdot g(p,q), \\
\boldsymbol{Y}_{LH}(i,j) &= \sum_{p,q} \boldsymbol{F}(i+p, j+q) \cdot h_{\text{LH}}(p,q), \\
\boldsymbol{Y}_{HL}(i,j) &= \sum_{p,q} \boldsymbol{F}(i+p, j+q) \cdot h_{\text{HL}}(p,q), \\
\boldsymbol{Y}_{HH}(i,j) &= \sum_{p,q} \boldsymbol{F}(i+p, j+q) \cdot h_{\text{HH}}(p,q).
\end{aligned} \tag{7}$$

Among them, $\boldsymbol{Y}_{LL}(i,j), \boldsymbol{Y}_{LH}(i,j), \boldsymbol{Y}_{HL}(i,j)$ and $\boldsymbol{Y}_{HH}(i,j)$ are the four wavelet coefficients after the convolution operation, while $g(p,q)$, $h_{\text{LH}}(p,q)$, $h_{\text{HL}}(p,q)$ and $h_{\text{HH}}(p,q)$ are the corresponding convolution filters. $\boldsymbol{F}$ denotes the input feature map, i and j are the output position indices, and p, q are the filter kernel positions. These convolution operations perform weighted sum calculations on each wavelet coefficient to further extract multi-scale features and enhance the model's ability to capture changes in different frequencies and directions.

These enhanced features are then merged back through the inverse discrete wavelet transform (IDWT) to obtain the fused features:

$$\boldsymbol{F}_{\text{WCP}} = \text{IDWT}(\boldsymbol{F}_{\text{LL}}, \boldsymbol{F}_{\text{LH}}, \boldsymbol{F}_{\text{HL}}, \boldsymbol{F}_{\text{HH}}), \tag{8}$$

this step ensures that the wavelet transform enhances the input features while maintaining multi-scale information.

3) Multi-Scale Pooling Mechanism: In order to further extract important information with features of different scales, we designed a multi-scale pooling mechanism (MPM). This mechanism aggregates features by applying pooling filters of various sizes ($3 \times 3, 5 \times 5, 7 \times 7$). The pooling process helps capture speech variation features at several scales, especially for multi-scale features of audio deepfake. The specific pooling process is as

$$\begin{aligned}
\boldsymbol{F}_{pool_{3\times3}} &= \text{MaxPool}(\boldsymbol{F}_{WCP}, \text{kernel}_{\text{size}} = 3), \\
\boldsymbol{F}_{pool_{5\times5}} &= \text{MaxPool}(\boldsymbol{F}_{WCP}, \text{kernel}_{\text{size}} = 5), \\
\boldsymbol{F}_{pool_{7\times7}} &= \text{MaxPool}(\boldsymbol{F}_{WCP}, \text{kernel}_{\text{size}} = 7).
\end{aligned} \tag{9}$$

Then, the three pooling results are averaged and pooled to obtain the enhanced feature representation of multi-scale pooling:

$$\boldsymbol{F}_{MPM} = \frac{\boldsymbol{F}_{pool_{3\times3}} + \boldsymbol{F}_{pool_{5\times5}} + \boldsymbol{F}_{pool_{7\times7}}}{3}, \tag{10}$$

this step can effectively capture the different time scale characteristics of speech signals and has strong recognition capabilities for artifacts.

TABLE I
SUMMARY OF THE DATASET AND DETAILED STATISTICS

Datasets	Genuine	Spoofed	Total
Train	2,580	22,800	25,380
Dev	2,548	22,296	24,844
Eval(19LA)	7,355	63,882	71,237
Eval(21LA)	18,452	163,114	181,566
Eval(21DF)	22,617	589,212	611,829
Eval(FoR)	2,264	2,370	4,634
Eval(ITW)	19,963	11,816	31,779
Eval(15LAE)	9,404	184,000	193,404

D. Gated Feed Forward Network

To increase feature expressiveness in gated feed forward network (GFN), we use gating methods and convolutional layers. The function of GFN is to convert the pooled features and send them on to the next MWPM. To accomplish this, we first perform convolution operations to the pooled features, then weight them using a gating mechanism, which is based on the Gelu activation function. In this approach, GFN can efficiently manage the flow of key features while minimizing the interference of useless information. Fig. 1 depicts the GFN design, which helps manage the flow of critical features and fine information to the next level of the MWPM block. The specific formula process is as follows. First, we use convolution on the pooled features:

$$\boldsymbol{F}_{Conv} = \text{Conv}(\boldsymbol{F}_{MPM}), \tag{11}$$

next, the gating mechanism is applied to control the flow of features:

$$\boldsymbol{F}_{gate} = \text{Conv}(\text{Gelu}(\boldsymbol{F}_{Conv}) \times \boldsymbol{F}_{Conv}), \tag{12}$$

the operator Conv denotes a one-dimensional convolution applied along the feature dimension, which enables the model to refine and transform the learned representations. The Gelu function refers to the gaussian error linear unit activation, which introduces smooth non-linearity to enhance representation learning. This approach ensures that useful features are effectively propagated in the network while avoiding the influence of irrelevant information.

Finally, a fully connected layer with two output nodes is employed to perform binary classification. Since the preceding modules effectively capture rich and discriminative representations, this simple linear classifier is sufficient to achieve strong performance without introducing further architectural complexity.

IV. EXPERIMENTS

A. Datasets

To verify the efficiency and generalization ability of our proposed method, we conducted experiments on multiple public datasets. The detailed statistics of the datasets used in the experiment are shown in Table I. The model was trained on the ASVspoof 2019 logical access (19LA) [43] training set. This dataset is a widely benchmark for audio deepfake detection. During testing, we assessed performance on diverse datasets, including 19LA test set, ASVspoof 2021 logical access (21LA) and deepfake (21DF) [11], Fake-or-Real (FoR) [44], In-the-Wild (ITW) [45], and ASVspoof 2015 logical access evaluation (15LAE) [46]. The 21LA dataset includes genuine and forged voices transmitted through telephone systems, testing robustness in diverse conditions. The 21DF dataset, with over 100 deception techniques and lossy compression formats, demands strong generalization. The FoR dataset integrates diverse speech sources, while ITW collects real-world audio forgery data from social media. The 15LAE dataset serves as an early baseline for evaluating forgery detection models.

B. Metrics

To comprehensively assess the model's effectiveness and its impact on the Automatic Speaker Verification (ASV) system, this study employs two primary evaluation metrics, the equal error rate (EER) and the minimum tandem detection cost function (min t-DCF). For the 19LA and 21LA datasets, both EER and min t-DCF are used to evaluate the model's detection performance and its interaction with ASV systems. In the deepfake detection challenge, the EER is adopted as the primary metric to measure the model's ability to distinguish between genuine and forged speech. Consequently, for the 21DF, FoR, ITW, and 15LAE datasets, only EER is used to evaluate performance.

1) EER: The EER is a key metric determined by two factors, the false acceptance rate (P_{FAR}) and the false rejection rate (P_{FRR}). P_{FAR} represents the probability of mis-classifying synthesized speech samples as genuine when their detection score exceeds a predefined threshold θ. Conversely, P_{FRR} quantifies the probability of mis-classifying genuine speech samples as fake when their detection score falls below the threshold θ. The calculations for P_{FAR} and P_{FRR} are as follows:

$$P_{\text{FAR}}(\theta) = \frac{\text{spoofed trials with score} > \theta}{\text{total spoofed trials}}, \tag{13}$$

$$P_{\text{FRR}}(\theta) = \frac{\text{genuine trials with score} \leq \theta}{\text{total genuine trials}}. \tag{14}$$

P_{FAR} evaluates the model's ability to avoid false positives, with lower values indicating higher accuracy in identifying synthesized speech. Similarly, P_{FRR} assesses the model's capability to avoid false negatives, with lower values signifying better performance in recognizing genuine speech. As θ increases, P_{FAR} decreases while P_{FRR} increases. The EER value is defined at the point where $\text{EER} = P_{FAR}(\theta_{EER}) = P_{FRR}(\theta_{EER})$. A lower EER indicates better overall detection performance, making it a necessary metric for evaluating the effectiveness of deepfake detection models.

2) min t-DCF: The min t-DCF is not only used to evaluate the performance of spoofed speech detection systems but, more importantly, to assess the impact of spoofing attacks and defenses on the reliability of ASV systems in practical scenarios. The min t-DCF calculation accounts for the ASV system's susceptibility to spoofing attacks and its ability to distinguish between genuine and spoofed speech. The calculation formula is as follows:

$$\min \text{t-DCF} = \min_{\theta} \{\beta P_{\text{FRR}}(\theta) + P_{\text{FAR}}(\theta)\} \tag{15}$$

For a given spoofing algorithm, the weight factor β is inversely proportional to the ASV system's FAR for that specific algorithm. If a spoofing algorithm severely disrupts the ASV system's classification performance, the penalty imposed by the metric will be higher. This design ensures that min t-DCF not only evaluates the detection system's accuracy but also emphasizes its robustness and reliability in mitigating real-world spoofing threats.

C. Comparative Models for Audio Deepfake Detection

- TSSDN [47]: This lightweight neural network integrates ResNet-style skip connections with Inception-style parallel convolutions. This hybrid design achieves a commendable balance between computational efficiency and detection performance.
- RawGAT-ST [48]: This model directly learns feature representations from raw waveform inputs and incorporates feature fusion at the model level. By combining spectral and temporal sub-graphs and employing a graph pooling method, it significantly enhances the discrimination between genuine and forged audio.
- AASIST [19]: A novel architecture based on heterogeneous stacked graph attention layers. It models forged speech features across temporal and spectral intervals using heterogeneous attention mechanisms and stacked node operations to achieve superior detection accuracy.
- AASIST-L [19]: A lightweight version of AASIST, designed to reduce computational overhead while maintaining comparable detection performance, making it suitable for resource-constrained environments.
- LC-Res18 [49]: This dual-branch network uses LFCC and CQT as inputs, employing multi-task learning and a convolutional block attention module to enhance the representation of forged features. It is particularly effective in detecting various types of voice forgeries.
- W2V2-ST [50]: This approach combines W2V2 as a front-end feature extractor with AASIST as a back-end classifier. It leverages advanced data augmentation techniques to achieve substantial improvements in forged speech detection performance. Since neither our method nor the comparison methods use data augmentation, we excluded it from our experiment for a fair comparison.
- WL-Res18 [51]: A dual-attention network that integrates W2V2 and Log-mel features via self-attention mechanisms. It incorporates a enhanced ResNet architecture to improve the accuracy of forged speech detection.
- AMSDF [52]: It detects audio forgeries by leveraging multi-view correlations among audio, emotion, and text. It uses an intra-view graph attention mechanism for intra-view feature aggregation, a heterogeneous graph fusion module for cross-view correlation, and a group-based readout scheme to retain differentiating features, achieving efficient distinction between real audio and forged audio.

D. Experimental Details

During the training phase, all models were trained in batch size of 8 and optimized utilizing the Adam optimizer. The optimizer was set to the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$ and weight decay of 10^{-4}. The optimization approach followed the weighted cross-entropy loss function, with a learning rate of 1×10^{-6} . To ensure model convergence and robustness, 100 epochs of training were run. The entire framework was developed using Python 3.8 with the PyTorch 1.8 deep learning library. The wavelet transform operations were implemented using the pytorch_wavelets library. All experiments were conducted on a Linux operating system equipped with an NVIDIA GeForce RTX 4090 GPU and 128 GB of RAM. CUDA version 12.2 was used for GPU acceleration. In the experiment, we chose six multi-scale wavelet pooling modules (MWPMs), which are further validated in the subsequent analysis.

V. Experiments Results Analysis

A. Comparative Experiments

1) Results on ASVspoof 2019 LA Dataset: To comprehensively evaluate the performance of our proposed method in the forged speech detection task, we conducted comparative experiments on the ASVspoof 2019 LA evaluation set, and the experimental results are shown in Table II. Compared to the baseline end-to-end TSSDN, the spectro-temporal graph attention network RawGAT-ST achieved significant performance improvements, with an EER reduction of 0.55% and a min t-DCF reduction of 0.0135. These improvements are primarily attributed to its ability to utilize spectral-temporal graph attention mechanisms, which effectively capture correlations across time and frequency domains in forged speech. AASIST, as an innovative GNN, demonstrated outstanding performance in forgery detection, achieving an EER of 0.82% and a min t-DCF of 0.0272 on the 19LA dataset, representing a considerable enhancement compared to earlier approaches. Its lightweight variant, AASIST-L, maintained robust detection performance, highlighting the advantages of efficient and lightweight model design. LC-Res18, which employs a combination of LFCC and CQT as inputs, achieved high detection accuracy by enhancing forged feature representation through multi-task learning. Despite using ResNet18 as its back-end classifier, LC-Res18 outperformed more complex models like AASIST on the ASVspoof 2019 LA evaluation set, achieving an EER of 0.80% and a min t-DCF of 0.0210. This result suggests that combining diverse speech features could enhance detection performance. W2V2-ST, by fine-tuning the pre-trained W2V2 model and integrating it with the AASIST back-end classification network, achieved remarkable progress in detecting forged speech. It recorded an EER of 0.25% and a min t-DCF of 0.0071 on the 19LA evaluation set, reducing the EER by 0.55% compared to previous SOTA method. Furthermore, AMSDF, through the integration of multi-view features encompassing audio, emotion, and text, outperformed other detection approaches, demonstrating the effectiveness of multi-view feature fusion in handling the challenges of forged speech detection.

TABLE II
COMPARISON OF DETECTION PERFORMANCE BASED ON EER(%) AND MIN T-DCF ON THE ASVSPOOF 2019 LA EVALUATION SET(19LA). A07–A19 DENOTE 13 SPOOFING ALGORITHMS USED IN THE EVALUATION SET, INCLUDING 11 UNKNOWN ATTACKS AND 2 KNOWN ONES. OUR RESULTS ARE HIGHLIGHTED WITH GRAY SHADING AND BEST EXPERIMENTAL RESULTS ARE SHOWN IN BOLD

Method	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	EER	min t-DCF
TSSDN [48]	1.43	0.75	0.02	1.75	**0.06**	0.18	0.06	0.11	2.05	1.25	6.01	1.14	1.44	1.62	0.0474
RawGAT-ST [49]	0.34	0.53	0.08	0.38	0.29	0.33	0.30	0.29	0.27	0.83	1.71	3.11	0.73	1.07	0.0339
AASIST [19]	0.47	0.41	0.00	0.79	0.16	0.67	0.12	0.16	0.51	0.65	1.28	2.65	0.67	0.82	0.0272
AASIST-L [19]	0.61	0.20	0.00	1.06	0.16	0.77	0.20	0.06	0.61	0.65	1.91	3.03	0.69	0.99	0.0317
LC-Res18 [50]	0.05	0.26	0.00	**0.26**	0.13	0.17	0.18	0.10	0.18	0.10	2.48	**0.40**	**0.17**	0.80	0.0210
W2V2-ST [51]	0.06	0.06	0.02	0.40	0.10	0.14	0.00	0.06	0.24	**0.06**	0.37	0.84	0.35	0.25	0.0071
AMSDF	0.03	0.05	0.03	0.52	0.38	0.23	0.02	0.05	0.33	0.12	0.30	0.47	0.41	0.31	0.0097
AWaveformer	**0.02**	**0.00**	**0.00**	0.40	0.12	**0.09**	**0.00**	**0.04**	**0.10**	0.09	**0.26**	0.42	0.30	**0.13**	**0.0042**

TABLE III
COMPARISON OF ROBUSTNESS AND EFFICIENCY BASED ON EER(%) AND MIN T-DCF ON THE ASVSPOOF 2021 LA DATASET(21LA), AND EER(%) ON THE ASVSPOOF 2021 DF DATASETS(21DF)

Methods	21LA		21DF
	EER	min t-DCF	EER
TSSDN [44]	10.17	0.4627	23.22
RawGAT-ST [49]	10.91	0.4931	20.96
AASIST [19]	12.82	0.5468	20.25
AASIST-L [19]	13.86	0.5743	29.98
W2V2-ST [51]	7.36	0.3769	8.13
WL-Res18 [52]	4.12	0.3008	5.34
AMSDF	3.63	0.2859	6.06
AWaveformer	**2.33**	**0.2486**	**3.63**

TABLE IV
COMPARISON OF GENERALIZATION PERFORMANCE BASED ON EER(%) ACROSS THE IN-THE-WILD(ITW), FAKE-OR-REAL(FOR), AND ASVSPOOF 2015 LOGICAL ACCESS EVALUATION DATASETS(15LAE)

Method	ITW	FOR	15LAE
TSSDN [48]	34.16	36.17	10.58
AASIST [19]	43.50	44.26	3.22
AASIST-L [19]	44.64	31.40	5.25
W2V2-ST [51]	18.67	8.98	0.26
AMSDF	18.06	8.30	**0.13**
AWaveFormer	**10.25**	**5.15**	0.21

Unlike AMSDF and W2V2-ST, which primarily rely on W2V2 for extracting self-supervised pre-trained features, our method integrates WavLM and merges the features of both W2V2 and WavLM through a cross-attention mechanism. To further help the model capture artifacts and reduce time complexity, we replace the conventional token mixer in the Transformer with wavelet transform and multi-scale pooling. Our method achieves an EER of 0.13% and a min t-DCF of 0.0042 on the 19LA dataset, resulting in a 0.12% reduction in EER compared to the previous best result, demonstrating the development and effectiveness of our approach.

2) Results on ASVspoof 2021 LA and DF Datasets: The challenge of detecting forgeries in the 21LA and 21DF datasets increases significantly due to the impact of transmission and interference noise on the voice stream. The experimental results are presented in Table III. The EER of classic end-to-end models, such as TSSDN, RawGAT-ST, and AASIST, on the 21DF dataset exceeds 20%, highlighting the limitations of these methods in handling complex noise environments. W2V2-ST improves forgery detection performance by utilizing self-supervised features, achieving an EER 7.92% lower than the previous SOTA technique. This indicates that self-supervised features can substantially enhance forgery detection capabilities in challenging scenarios. WL-Res18 also employs the self-attention mechanism to combine W2V2-extracted features with handcrafted Logmel features, resulting in richer and more robust speech representation. WL-Res18 achieves superior performance on the 21DF dataset, obtaining an EER of 5.34%, while achieving 4.12% EER and a min t-DCF of 0.3008 on the 21LA dataset. Furthermore, AMSDF demonstrates exceptional performance in forgery detection by integrating features from three different perspectives, text, emotion, and audio, while employing a well-designed GNN as the back-end classifier. This approach highlights the effectiveness of multi-view feature fusion, as shown by the reduction in EER and t-DCF by 0.49% and 0.0149, respectively, on the 21LA dataset, compared to WL-Res18.

In comparison to previous methods, our approach integrates two self-supervised pre-trained features, W2V2 and WavLM, to obtain more comprehensive feature representations. By leveraging a cross-attention mechanism, this method enables deep fusion of the pre-trained features, while optimizing the token mixer through wavelet transform and multi-scale pooling. These improvements lead to significant gains in detecting forged speech. Specifically, our method achieves an EER of 2.33% and a min t-DCF of 0.2486 on the 21LA dataset, and an EER of 3.63% on the 21DF dataset. These results outperform existing techniques, demonstrating that our approach not only enhances model detection capabilities but also maintains strong performance in challenging noisy conditions, highlighting its robustness and effectiveness.

3) Results on ITW, for and 15LA Datasets: To further evaluate the generalization capability of the proposed method, we conducted cross-database experiments on the ITW, FOR and 15LAE datasets. The experimental results, summarized in Table IV, reveal that traditional detection frameworks without pre-trained models, such as AASIST, AASIST-L, and TSSDN, show limited performance across these datasets. This indicates significant deficiencies in generalization capabilities in cross-domain environments. In contrast, incorporating self-supervised learning, particularly with W2V2, significantly enhances forgery detection performance. The W2V2-based framework, W2V2-ST, achieved EERs of 18.67%, 8.98%, and 0.26% on the ITW, FOR, and 15LAE datasets, respectively. These results demonstrate substantial improvement, with EERs reduced by over 10%

TABLE V
THE ABLATION STUDY AIMS TO DEMONSTRATE THE EFFECTIVENESS OF EACH COMPONENT OF DETECTION SYSTEM. W2V2: WAV2VEC 2.0 PRE-TRAINED MODEL FEATURE EXTRACTION. WAVLM: WAVLM PRE-TRAINED MODEL FEATURE EXTRACTION. A-F: CROSS-ATTENTION FUSION. P-E: POSITIONAL ENCODING. WCP: WAVELET COEFFICIENT PROCESSING. MPM: MULTI-SCALE POOLING MECHANISM. GFN: GATED FEED FORWARD NETWORK

Method	21LA	21DF	ITW	FOR	15LAE
Model I (w/o W2V2)	18.45	46.21	48.56	48.23	39.67
Model II (w/o WavLM)	**1.52**	4.72	40.08	40.67	20.41
Model III (w/o A-F)	4.52	5.37	11.97	10.12	0.16
Model IV (w/o P-E)	7.84	5.95	12.11	10.72	0.14
Model V (w/o WCP)	2.62	6.93	15.59	10.42	0.21
Model VI (w/o MPM)	6.36	5.31	**7.58**	6.04	0.28
Model VII (w/o GFN)	4.84	4.93	14.30	11.13	**0.11**
AWaveFormer	2.33	**3.63**	10.25	**5.15**	0.21

compared to traditional methods, highlighting the effectiveness of self-supervised features in cross-domain detection tasks. Additionally, multi-view feature fusion has proven to be pivotal in improving detection accuracy. Specifically, AMSDF combines W2V2-extracted features with emotional and RawNet [53] speech features, processed through the GNN, yielding better results. In comparison to W2V2-ST, their method reduced the EER by 0.61%, 0.68%, and 0.13% on the ITW, FOR, and 15LAE datasets, respectively, further emphasizing the positive impact of multi-view feature fusion.

Building on this, the method proposed in this paper fully integrates W2V2 and WavLM via a cross-attention mechanism, optimizing the model structure with wavelet transform, multi-scale pooling, and novel gated feed forward network. The resulting model demonstrates superior detection performance on the ITW and FOR datasets, achieving EERs of 10.25% and 5.15%, respectively. When compared to the AMSDF, our proposed method outperforms with reductions of 7.81% and 3.15%, respectively, showcasing its exceptional cross-domain detection capabilities. Although the detection performance on the 15LAE dataset is slightly lower than that of some other approaches, it remains competitive. These experimental results affirm that the proposed AWaveFormer not only improves forgery detection performance but also exhibits robust generalization ability across different databases, validating the efficacy and robustness of the model architecture in audio deepfake detection.

B. Ablation Experiment

To further validate the effectiveness of the proposed approach, we conducted ablation studies to analyze the individual contributions of WavLM, W2V2, cross-attention fusion, position embedding, wavelet coefficient processing (WCP), multi-scale pooling mechanism (MPM), and gated feed forward network (GFN) in the AWaveFormer model. The experiments were carried out on several public datasets, including 21LA, 21DF, ITW, FoR, and 15LAE.

Keeping the back-end model of the multi-scale wavelet Transformer network unchanged, we first evaluate the impact of the front-end feature extraction module. Model I only uses WavLM as the input feature extractor. As shown in the first row of Table V , the model's performance across all five datasets is suboptimal, indicating that WavLM alone is not well-suited for the audio deepfake detection task. In the following phase, we substituted WavLM with W2V2 as the front-end feature extractor to construct Model II. While Model II performed well on 21LA and 21DF, with EERs of 1.52% and 4.72%, respectively, its performance on the other three datasets was unsatisfactory, indicating its low generalization capacity. As a result, it is challenging for a single pre-trained feature to maintain consistently strong performance across all evaluation scenarios. In contrast, after fusing the features from both W2V2 and WavLM, the proposed method achieves competitive performance across all test sets. This demonstrates that these two models provide complementary information, and their integration enhances the model's overall detection capability and generalization.

To investigate the role of the cross-attention mechanism, we simply fused the features from WavLM and W2V2 and input them into the multi-scale wavelet Transformer network, denoted as Model III. This approach significantly improved performance across all five datasets. Model III exhibits a strong detection effect, with EER reductions of nearly 30% on the ITW, FoR, and 15LAE datasets compared to using WavLM or W2V2 alone. These results indicate that both WavLM and W2V2 are important for detecting fake speech and that their joint use enhances detection efficiency. However, simple feature fusion does not fully leverage the potential of pre-trained models. The AWaveFormer further improved performance, reducing the EER on 21LA, 21DF, ITW, and FoR by 2.19%, 1.74%, 1.72%, and 4.97%, respectively, compared to Model III. This highlights the importance of the cross-attention fusion mechanism in fully utilizing the speech features.

Next, we examined the effects of WCP, MPM, GFN, and position embedding within the MWTN. In these trials, the front-end feature extraction remained consistent with AWaveFormer, using cross-fused WavLM and W2V2 features. Model IV, which omits position embedding, exhibits a performance drop, demonstrating that positional encoding helps the model capture the temporal relationships in the input sequence, essential for understanding the temporal patterns of fake speech. The model's generalization ability is also impaired when the WCP module is removed, as demonstrated by Model V, which yields EERs of 15.59% and 10.42% on the ITW and FoR datasets, respectively. These results confirm that WCP is essential for detecting forged speech. When the MPM is removed, the detection performance declines obviously, as shown in Model VI, underscoring the critical role of the MPM in the model. Moreover, when the GFN module is removed, as illustrated in Model VII, there is a obvious decline in performance, further highlighting the importance of each module within AWaveFormer. These results collectively demonstrate that the different components of AWaveFormer function synergistically, contributing to its outstanding performance in forgery detection.

These results reveal that while all components contribute to the overall performance, the removal of the WCP and position embedding modules have the most significant negative impact on detection performance, highlighting their critical roles in the success of AWaveFormer.

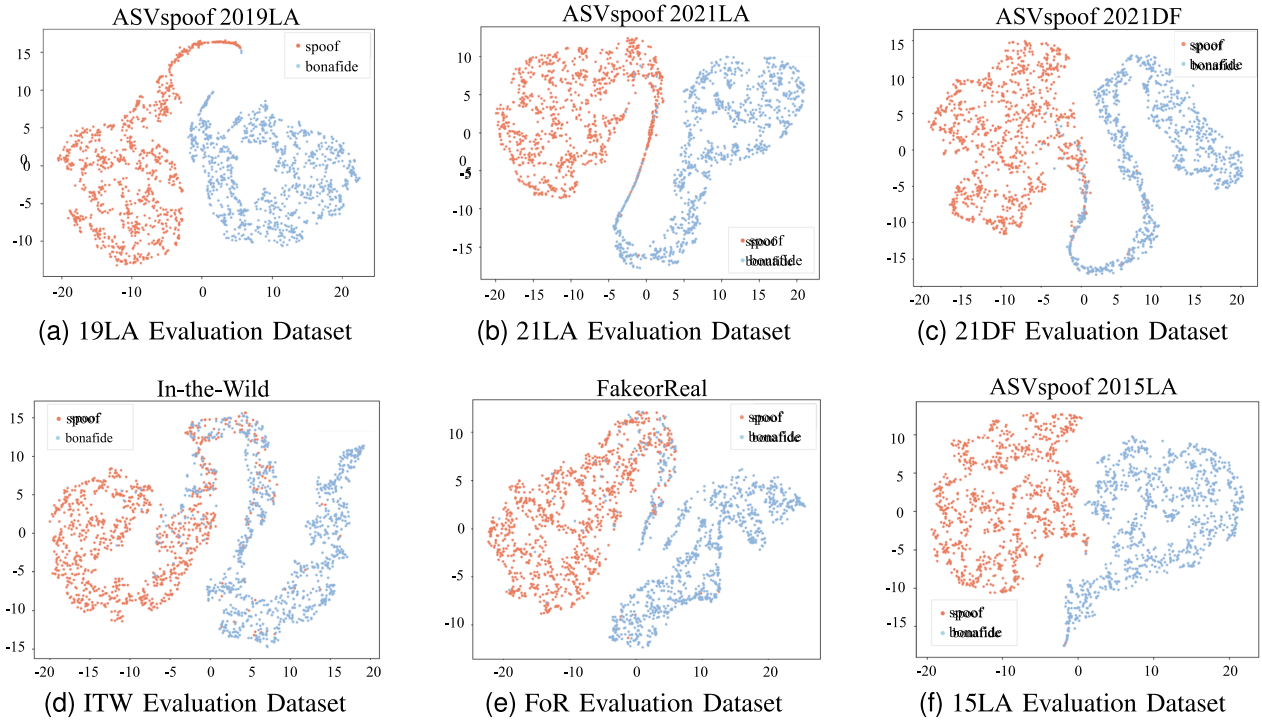

Fig. 5. t-SNE plot of AWaveFormer on ASVspoof 2019 LA evaluation dataset (a), ASVspoof 2021 LA dataset (b), ASVspoof 2021 DF dataset (c), In-the-Wild dataset (d), Fake-or-Real dataset (e), ASVspoof 2015 LA evaluation dataset (f). The t-SNE plot was generated by randomly selecting 1000 samples from each class. Bonafide samples are marked in blue, while spoof samples are highlighted in orange.

TABLE VI
PERFORMANCE COMPARISON OF MAINSTREAM BACKENDS FOR AUDIO DEEPFAKE DETECTION. RES18: RESNET 18. GRIFFIN: PROPOSED BY [54]. TRANSFORMER: TRADITIONAL TRANSFORMER NETWORK, HERE WE ONLY USE THE ENCODER PART. AASIST: PROPOSED BY [19]

Method	21LA	21DF	ITW	FOR	15LAE
Model I (res18)	**2.18**	5.94	13.76	13.83	0.46
Model II (Griffin)	8.25	6.64	13.05	16.16	0.21
Model III (Transformer)	3.02	4.32	**10.10**	16.37	0.28
Model IV (AASIST)	6.72	3.71	10.41	11.39	**0.20**
Ours	2.33	**3.63**	10.25	**5.15**	0.21

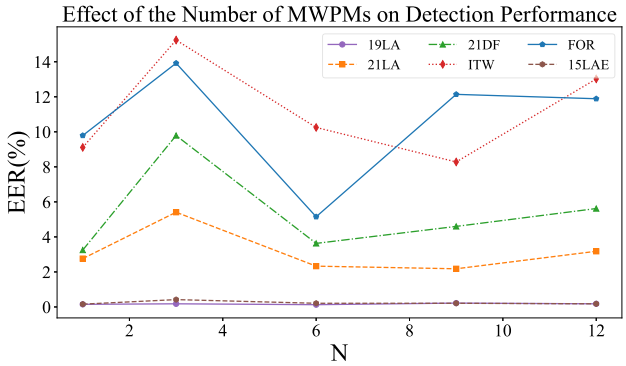

Fig. 6. Impact of the number of multi-scale wavelet pooling modules on detection performance (EER%).

C. Comparison of Mainstream Backend Networks

In the comparative experiment replacing back-end models, we evaluated several prominent back-end architectures, including ResNet18, the Griffin network recently proposed by Google DeepMind [54], the classic Transformer model, and AASIST. These models are referred to as Model I, Model II, Model III, and Model IV, respectively. The experimental results, shown in Table VI, were obtained using the 21LA, 21DF, ITW, FoR , and 15LAE datasets.

Model I employs ResNet18 as the back-end and achieves an EER of 2.18% on the 21LA dataset, which is relatively satisfactory. However, its performance on the other datasets is unsatisfactory, suggesting that simple CNN struggles to process and leverage complex pre-trained features. Model II uses the Griffin network, which combines gated linear recursion and local attention techniques, excelling in speech modeling. Despite its strength in speech modeling, its performance in speech forgery detection remains limited, failing to improve detection capabilities. Model III employs the classic Transformer model as the back-end, yielding EERs of 3.02%, 4.32%, 10.10%, 16.37%, and 0.28% on the 21LA, 21DF, ITW, FoR, and 15LAE datasets, respectively. These results demonstrate the potential of the Transformer model in speech forgery detection, offering valuable insights and a new direction for future research in this field. Finally, Model IV uses AASIST, a novel graph neural network. It demonstrates solid performance across all five datasets, showcasing the effectiveness of graph neural networks in forgery detection tasks. While AASIST performs admirably, it still leaves room for improvement compared to our proposed AWaveFormer model. Our model outperforms

conventional convolutional networks, traditional Transformers, recurrent neural networks, and graph neural networks in terms of detection capability, owing to its integration of position embedding, wavelet transform, multi-scale pooling, and the gated feed forward network. In conclusion, our AWaveFormer model outperforms the conventional models across multiple experiments, demonstrating its superior ability to detect synthesized speech.

D. Further Analysis

1) T-SNE Visualization: To further demonstrate the classification effect of AWaveFormer, we plotted t-SNE visualizations on the 19LA, 21LA, 21DF, ITW, FOR, and 15LAE datasets, as shown in Fig. 5. To conduct the analysis, we randomly picked 1,000 true and fake samples from each dataset. The visualization results show that on the 19LA, FoR, and 15LAE datasets, true and fake samples are almost completely separated, and samples of the same category are clustered together, reflecting the excellent classification performance of the model on these datasets. However, on the 21LA, 21DF, and ITW datasets, although some true and fake samples are mixed together and difficult to completely distinguish, the model can still distinguish true and fake samples well overall. This further verifies the strong classification ability of AWaveFormer on different datasets, indicating that it can maintain good performance in a variety of audio deepfake detection tasks.

2) Effect of MWPM Number on Detection Performance: We further explored the effect of the number of MWPMs on the performance of forged speech detection. Specifically, we selected five different values of N: 1, 3, 6, 9, and 12, and tested them on the 19LA, 21LA, 21DF, ITW, FoR, and 15LAE datasets. The experimental results are depicted in Fig. 6. On these six datasets, in most cases, when $N = 6$, the detection performance of the model is optimal. This demonstrates that when the number of MWPMs is too small, the model may struggle to fully understand the self-supervised features, and when the number is too large, the complexity of the model increases, potentially leading to a drop in detection performance. After comprehensive consideration, we selected 6 MWPMs, and the experimental results verified that the effect of forged speech detection is best when $N = 6$.

VI. Conclusion

In this work, we propose a novel audio deepfake detection architecture based on Transformer network, named AWaveFormer. The proposed framework consists of three key components: pre-trained feature extraction, cross-attention feature fusion, and the multi-scale wavelet Transformer network (MWTN) for back-end classification. First, we extract features using two pre-trained models, Wav2Vec 2.0 and WavLM. The Wav2Vec 2.0, which was pre-trained on large-scale corpora, making it particularly effective for speech feature extraction. WavLM, on the other hand, not only learns ASR tasks through masked speech prediction but also captures knowledge from non-ASR tasks through speech denoising modeling, providing strong cross-task capabilities. These two self-supervised models have been extensively trained and are adept at extracting comprehensive and robust speech features. To maximize the utility of these pre-trained features, we employ a cross-attention mechanism to fuse the features from Wav2Vec 2.0 and WavLM, allowing the model to extract more robust and discriminative features. Subsequently, we replace the token mixer in the Transformer architecture with wavelet transform and multi-scale pooling operations, which are designed to enhance feature extraction from multiple scales, resolutions, and improve efficiency. Experimental results demonstrate that the proposed method achieves impressive detection performance, particularly in real-world scenarios, showcasing its competitiveness and robustness.

References

[1] Z. Yin et al., "Improving deepfake detection generalization by invariant risk minimization," *IEEE Trans. Multimedia*, vol. 26, pp. 6785–6798, 2024.

[2] L. A. Passos et al., "A review of deep learning-based approaches for deepfake content detection," *Expert Syst.*, vol. 41, no. 8, 2024, Art. no. e13570.

[3] M. Kang, D. Min, and S. J. Hwang, "Grad-StyleSpeech: Any-speaker adaptive text-to-speech synthesis with diffusion models," in *Proc. 2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[4] A. Triantafyllopoulos et al., "An overview of affective speech synthesis and conversion in the deep learning era," *Proc. IEEE*, vol. 111, no. 10, pp. 1355–1381, Oct. 2023.

[5] S. Xiao, G. Lan, J. Yang, W. Lu, Q. Meng, and X. Gao, "MCS-GAN: A different understanding for generalization of deep forgery detection," *IEEE Trans. Multimedia*, vol. 26, pp. 1333–1345, 2024.

[6] Y. Xie et al., "The codecfake dataset and countermeasures for the universally detection of deepfake audio," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 386–400, 2025.

[7] C. Wang, J. He, J. Yi, J. Tao, C. Y. Zhang, and X. Zhang, "Multi-scale permutation entropy for audio deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1406–1410.

[8] N. M. Müller, P. Sperl, and K. Böttinger, "Complex-valued neural networks for voice anti-spoofing," 2023, *arXiv:2308.11800*.

[9] K. S. Krishnan and K. S. Krishnan, "MFAAN: Unveiling audio deepfakes with a multi-feature authenticity network," in *Proc. 9th Int. Conf. Signal Process. Commun.*, 2023, pp. 585–590.

[10] W. H. Kang, J. Alam, and A. Fathan, "CRIM's system description for the ASVSPoof2021 challenge," in *Proc. ASVspoof Workshop*, Sep. 2021, pp. 100–106, doi: 10.21437/asvspoof.2021-16.

[11] X. Liu et al., "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2507–2522, 2023.

[12] B. Wang, Y. Tang, F. Wei, Z. Ba, and K. Ren, "FTDKD: Frequency-time domain knowledge distillation for low-quality compressed audio deepfake detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4905–4918, 2024.

[13] A. Fathan, J. Alam, and W. Kang, "MultiResolution decomposition analysis via wavelet transforms for audio deepfake detection," in *Proc. Int. Conf. Speech Comput.*, 2022, pp. 188–200.

[14] Y. Yang et al., "A robust audio deepfake detection system via multi-view feature," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 13131–13135.

[15] C. Liu, X. Xu, and F. Xiao, "ASSD: An AI-synthesized speech detection scheme using whisper feature and types classification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 542–556, 2025.

[16] L. Wang, L. Yu, Y. Zhang, and H. Xie, "Generalizable speech spoofing detection against silence trimming with data augmentation and multi-task meta-learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3296–3310, 2024.

[17] G. S. Kashyap et al., "Fooling the forgers: A multi-stage framework for audio deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[18] Z. Jin, L. Lang, and B. Leng, "Wave-spectrogram cross-modal aggregation for audio deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[19] J.-w. Jung et al., "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6367–6371.

[20] Z. Lei, H. Yan, C. Liu, Y. Zhou, and M. Ma, "GMM-ResNet2: Ensemble of group ResNet networks for synthetic speech detection," in *Proc. ICASSP 2024-2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 12101–12105.
[21] J. Huang, C. Du, X. Zhu, S. Ma, S. Nepal, and C. Xu, "Anti-compression contrastive facial forgery detection," *IEEE Trans. Multimedia*, vol. 26, pp. 6166–6177, 2024.
[22] C. Fan et al., "Dual-branch knowledge distillation for noise-robust synthetic speech detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2453–2466, 2024.
[23] A. Khan and K. M. Malik, "Securing voice biometrics: One-shot learning approach for audio deepfake detection," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2023, pp. 1–6.
[24] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10809–10819, doi: 10.1109/cvpr52688.2022.01055.
[25] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.
[27] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised XLS-R and SLS classifier," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 6765–6773.
[28] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 2808–2812.
[29] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised WAVLM and multi-fusion attentive classifier," in *Proc. ICASSP 2024-2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 12702–12706.
[30] L. Li, T. Lu, X. Ma, M. Yuan, and D. Wan, "Voice deepfake detection using the self-supervised pre-training model huBERT," *Appl. Sci.*, vol. 13, no. 14, 2023, Art. no. 8488.
[31] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[32] A. K. S. Yadav, Z. Xiang, K. Bhagtani, P. Bestagini, S. Tubaro, and E. J. Delp, "Compression robust synthetic speech detection using patched spectrogram transformer," 2024, *arXiv:2402.14205*.
[33] L. Cuccovillo, M. Gerhardt, and P. Aichroth, "Audio transformer for synthetic speech detection via formant magnitude and phase analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 4805–4809.
[34] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[35] H.-s. Shin, J. Heo, J.-h. Kim, C.-y. Lim, W. Kim, and H.-J. Yu, "Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 10581–10585.
[36] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "Safeear: Content privacy-preserving audio deepfake detection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2024, pp. 3585–3599.
[37] Y. Xie et al., "Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception," 2025, *arXiv:2504.06753*.
[38] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," in *Proc. IEEE 5th Int. Symp. Internet Sounds*, 2024, pp. 1–5.
[39] U. Zbezhkhovska and O. Khapilin, "Deepfake audio detection with sinc and wavelet filters in RawNet2," in *Proc. Int. Conf. Inf. Commun. Technol. Educ., Res., Ind. Appl.*, 2024, pp. 273–284.
[40] U. Avaiya, N. Badlani, R. Baadkar, and K. Talele, "Audio deepfake detection using deep scattering network," in *Proc. 9th Int. Conf. Commun. Electron. Syst.*, 2024, pp. 2063–2069.
[41] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Representations*, Nov. 2016.
[42] F. Xu, S. Mei, G. Zhang, N. Wang, and Q. Du, "Bridging CNN and transformer with cross attention fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5522214.
[43] A. Nautsch et al., "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 3, no. 2, pp. 252–265, Apr. 2021.
[44] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *Proc. 2019 Int. Conf. Speech Technol. Hum.- Comput. Dialogue*, 2019, pp. 1–10.
[45] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," 2022, *arXiv:2203.16263*.
[46] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, 2014, Art. no. 3750.
[47] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1265–1269, 2021.
[48] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," 2021, *arXiv:2107.12710*.
[49] K. Ma, Y. Feng, B. Chen, and G. Zhao, "End-to-end dual-branch network towards synthetic speech detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 359–363, 2023.
[50] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022, *arXiv:2202.12233*.
[51] B. Wang, Y. Ma, Y. Tang, R. Wang, and M. Zhang, "Enhancing synthesized speech detection with dual attention using features fusion," in *Proc. Int. Conf. Comput. Appl. Technol.*, 2023, pp. 104–109.
[52] J. Wu, Q. Yin, Z. Sheng, W. Lu, J. Huang, and B. Li, "Audio multi-view spoofing detection framework based on audio-text-emotion correlations," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 7133–7146, 2024.
[53] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6369–6373.
[54] M. Garg, D. Ghosh, and P. M. Pradhan, "Gestformer: Multiscale wavelet pooling transformer network for dynamic hand gesture recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 2473–2483.