# AURORA: Adaptive Audio–Video Multi-Scale Attention Fusion for Deepfake Detection

Jie Xu*, Shan Chang*, Hongzi Zhu†
*School of Computer Science and Technology, Donghua University, Shanghai, China
†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
jiexu@mail.dhu.edu.cn, changshan@dhu.edu.cn, hongzi@sjtu.edu.cn

***Abstract*—With the rapid advancement of generative forgery technologies, the detection of multi-modal deepfake audio-video content has become an urgent demand in cyber security and forensic analysis. However, detecting audio–video deepfakes remains challenging: forgery traces are often subtle, dispersed, and highly resolution-dependent; existing multimodal methods rely on simple concatenation or shallow interactions, leading to insufficient exploitation of cross-modal consistency. To address these issues, we propose a Cross-level Multi-modal Fusion (CLMF) framework that progressively integrates audio cues into visual representations through cross-level attention, adaptively enhancing complementary information while suppressing redundancy. In addition, we design an Adaptive Audio Feature Enhancement module (AAFE) to highlight subtle frequency-domain artifacts often masked by noise, and a Multi-scale Visual Feature Enhancement module (MVFE) to capture both local texture inconsistencies and global structural distortions. These components jointly achieve robust and consistent cross-modal alignment of forgery traces, leading to significant improvements in deepfake detection performance. On the FakeAVCeleb benchmark, AURORA achieves an accuracy (ACC) of 94.32% and an area under the curve (AUC) of 93.66%, demonstrating superior performance.**

***Index Terms*—Deepfake Detection, Audio-Video, Multi-modal Learning, Attention Fusion**

## I. Introduction

In recent decades, significant progress has been made in fields such as machine learning and deep learning, leading to increasingly sophisticated image and video forgery techniques capable of generating highly realistic videos that are indistinguishable from authentic ones. Such advances pose severe challenges to information security and social stability [1]. As a form of synthetic media, deepfake technology involves replacing one person's portrait with that of another, or synthesizing and forging facial appearances, voices, or expressions. Although this technology opens new possibilities for entertainment and education, it also introduces substantial risks, including privacy violations and even historical revisionism, raising serious ethical and security concerns.

The emergence of GANs [2] in 2014 significantly improved the realism of synthetic data and greatly reduced the difficulty of forgery. In recent years, diffusion model-based deepfake techniques have gained widespread adoption, advancing from single-modality synthesis to cross-modality generation [3]. Deep neural networks are trained on real datasets to capture the characteristics and actions of individuals under specific environmental conditions. The trained network is then applied to images of another person and enhanced using additional computer graphics techniques, integrating the new subject seamlessly into the original scene. Specifically, an encoder analyzes the similarity between the source and target faces. Upon extracting shared facial features, a decoder processes the encoded representation and reconstructs the output while preserving the original image's characteristics. As deepfake generation techniques grow increasingly sophisticated, robust detection methods have become essential to counter potential malicious applications.

Despite notable advancements in deepfake detection, current approaches still exhibit several critical limitations [4]. Most detection systems are constrained by either unimodal analysis or computationally intensive architectures, significantly hindering real-time performance [5]. Furthermore, effective integration of multi-modal features—particularly the synergistic combination of visual and audio cues—has not been adequately investigated, presenting substantial opportunities for enhancing detection accuracy and system robustness [6].

In existing multi-modal detection algorithms, these shortcomings manifest in three key aspects. First, in terms of audio feature extraction, most methods remain confined to time-domain analysis, overlooking the potential value of frequency-domain information. This omission results in less expressive audio features, making it difficult to effectively distinguish between genuine and fake audio. Second, in visual feature extraction, current approaches insufficiently explore high-dimensional topological features within video frames, failing to capture subtle texture and structural variations introduced during video forgery. Furthermore, in audio–visual feature fusion, most existing methods adopt simple concatenation or weighted-sum strategies, without fully considering the complementarity and correlations between audio and visual features at different levels. This leads to redundant and weakly consistent fused features, ultimately impairing the overall performance of detection algorithms. Therefore, how to effectively leverage audio frequency-domain information and visual high-dimensional topological features, while achieving efficient audio–visual feature fusion, remains a pressing challenge for current multi-modal deepfake detection algorithms.

To address these issues, we propose a novel network called AURORA. For the audio and video modalities, we design

Shan Chang is the corresponding author.

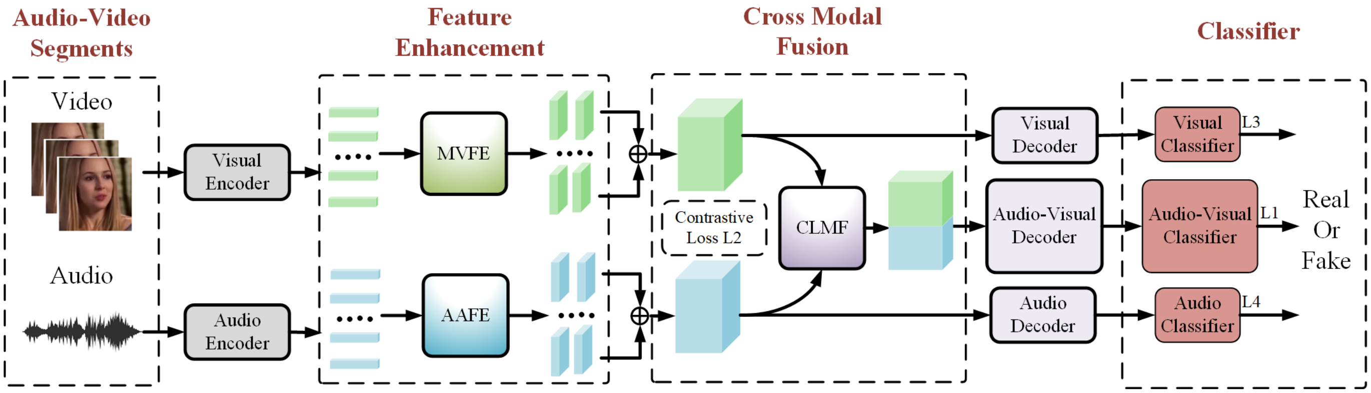

Fig. 1. Overview of the proposed network.

an adaptive audio feature module and a multi-scale visual feature module to enhance modality-specific features. For the audio–visual features, we introduce a cross-level multi-modal fusion module to achieve adaptive cross-level integration of audio and visual features. The main contributions of our work are as follows:

- For audio features, we propose an adaptive frequency-domain feature enhancement module that transforms raw audio representations from the time domain to the frequency domain via Fast Fourier Transform (FFT). An adaptive threshold-based segmentation strategy is then applied to effectively separate high- and low-frequency spectral components, which are subsequently individually enhanced and jointly fused using learnable weighting parameters. High-frequency components typically contain fine-grained temporal and spectral details, while low-frequency components reflect the global structural trend of the entire audio signal. By enhancing and adaptively fusing these complementary components, the discriminative representation power of audio features is significantly improved, making the differences between synthesized and genuine audio signals more apparent, thereby increasing the overall contribution of the audio modality to robust deepfake detection.
- For visual features, we design a dual-branch multi-scale aggregation enhancement network to extract video forgery-related high-dimensional topological representations and spatial correlations. Through channel-separated multi-scale aggregation operations, our approach effectively captures feature information across diverse spatial and temporal scales, thereby overcoming the intrinsic limitations of single-scale feature extraction mechanisms. This enables a more comprehensive and fine-grained representation of the video's authenticity cues and forgery artifacts, further enhancing the expressiveness, robustness, and discriminative ability of visual features and improving the overall accuracy of the visual modality in deepfake video detection tasks.
- For audio-video modality fusion, we propose a cross-level multi-modal fusion module that injects audio features into visual features using multi-level multi-modal fusion attention. By leveraging both self-attention and cross-attention mechanisms, the module constructs a cross-level enhancement structure that fuses features at different hierarchical levels. This fusion strategy not only accounts for the complementarity between audio and visual features but also dynamically adjusts the weights of features at different levels via the attention mechanism, reducing redundancy and preserving the consistent cross-modal components, thereby improving the effectiveness of the fused features.

We implement our approach on the FakeAVCeleb dataset and compare it with some multi-modal deepfake detection methods, including MRDF [7], AVoiD-DF [8], and AD-DFD [9]. Experimental results show that AURORA achieves 94.32% accuracy (ACC) and 93.66% area under the curve (AUC), outperforming MRDF by 0.27% ACC and 1.23% AUC.

## II. Related Work

Deepfake technology presents substantial risks to media information security, compromising confidentiality, integrity, and reliability. This growing threat has spurred significant research interest, leading to diverse detection methodologies in recent years. As categorized by Rana et al. [10], current detection approaches primarily fall into two paradigms: machine learning-based and deep learning-based methods.

**Machine Learning-based Method.** Machine learning models include Support Vector Machines (SVM), Logistic Regression, Multilayer Perceptrons (MLP), K-means clustering, Multiple Instance Learning (MIL), and Naive Bayes, among others. These methods use feature selection algorithms to construct feature vectors, which are then used as inputs to train classifiers for predicting whether audio or video content has been manipulated. Although machine learning models offer better interpretability and require less training time, the performance heavily depends on feature selection and extraction. [11] To enhance model robustness and accurately identify and distinguish relevant characteristics, deep learning algorithms have emerged as a more effective alternative.

**Deep Learning-based Method.** Deep learning models are now widely used in deepfake detection due to their inherent feature extraction and selection capabilities. These models can be further categorized into Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Region-based Convolutional Neural Networks (R-CNN), and Transformer architectures [12].

Recent studies have shifted from single-to-multiple models and single-modal to multi-modal approaches to improve deepfake detection performance. Abhishek et al. [13] proposed a hybrid deepfake detection framework that integrates CNNs and RNNs with spatiotemporal inconsistency detection capabilities. Chen et al. [14] partitioned facial regions within video frames and explored the relationships among them. By exploiting the correlations and temporal features between facial regions in deepfake videos, a feature graph is constructed where vertices represent region features extracted via CNN and edges represent inter-region correlations across the video [15]. A graph neural network is then applied to determine whether the video has been tampered with. Doan et al. [16] addressed the problem of arbitrary-content audio splicing in speech synthesis technologies by proposing BTS-E, a framework that evaluates the correlations among breathing, speech, and silence segments within audio clips, leveraging this information for deepfake detection tasks [17].

Sneha et al. [18] introduced an audio–visual deepfake detection method that combines fine-grained deepfake identification with binary classification. By integrating modality-specific labels [19], samples are categorized into four distinct types, thereby enhancing detection performance in both in-domain and cross-domain testing scenarios. Muhammad et al. [20] proposed MMDF-Net, a novel multi-modal framework for deepfake detection that integrates visual and audio features. MMDF-Net employs MobileNetV2 for lightweight local facial feature extraction and Swin Transformer to capture global facial relationships, ensuring comprehensive visual representation. For audio feature extraction, the Wav2vec model is adopted to complement visual data, enabling a holistic detection approach [21]. A multi-modal fusion mechanism is then used to integrate these features [22], facilitating an accurate and reliable real-time deepfake detection framework. Jin et al. [23] introduced a novel cross-modal deepfake audio detection framework that leverages multi-scale representations to improve generalization in distinguishing unseen synthetic utterances. Since spectrograms can reveal hidden characteristics within audio signals [24], the model learns discriminative and intrinsic feature representations by carefully aligning features from multiple modalities. Shirley et al. [25] combined video and audio features, leveraging the complementarity of different data types to detect inconsistencies introduced by deepfake. An attention mechanism is incorporated to focus on salient regions within each modality, further improving detection accuracy [26].

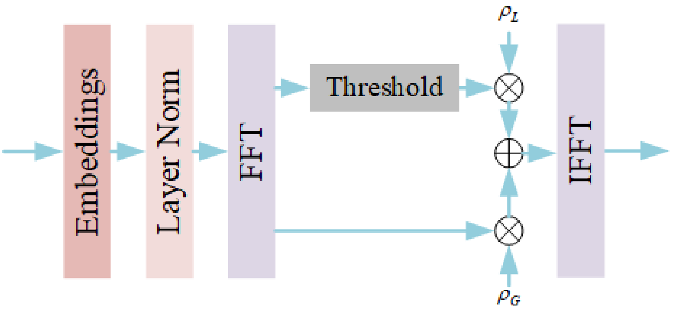

Fig. 2. Adaptive audio feature enhancement.

## III. Methodology

In this section, we present the proposed audio-video deepfake detection network in detail. As illustrated in Figure 1, different from MRDF [7], our network introduces three novel modules: (1) adaptive audio feature enhancement (AAFE); (2) multi-scale visual enhancement (MVFE); (3) an audio-visual feature fusion network based on cross-level multi-modal fusion (CLMF). By enhancing unimodal features and performing cross-level multi-modal fusion, the proposed network enables joint audio-video detection, effectively identifying various types of forgeries, including audio-only, video-only, and audio-video coordinated forgeries.

The proposed network accepts two types of input data: audio and video. Each modality is fed in batches to facilitate parallel processing and efficient feature extraction. The audio and video inputs are first processed by their respective frame-level feature extraction and enhancement networks, after which the enhanced representations are fused through a cross-layer multi-modal fusion module. The fused features are then passed to the classifier for final decision-making. As illustrated in the figure, the architecture includes multiple classifiers, comprising two single-modality classifiers and one multi-modal fusion classifier. All classifiers are implemented using fully connected layers.

The input audio signal is denoted as $X_{a,T_a}$, and the video signal is denoted as $X_{v,T_v}$. The number of input audio frames is $T_a$, and the number of input video frames is $T_v$. The target label for multi-modal deepfake detection is $y_{av}$, the target label for the audio single modality is $y_a$, and the target label for the video single modality is $y_v$.

### A. Adaptive Audio Feature Enhancement

For the audio feature extraction network, preliminary audio features $X_{a,t_a} \in \mathbb{R}^{C \times N}$ are first obtained using a pre-trained audio encoder, where $C$ denotes the number of signal channels and $N$ represents the length of the audio sample. As shown in Figure 2, the audio features are enhanced through processing by the time-frequency feature module.

We first divide the signal $x_{a,t_a}$ into a set of $M$ blocks $P_1, P_2, P_3, ...$, where each block has a predefined size $p$, i.e., each $P_i \in R^{C \times p}$. Each block is then mapped to a new dimension $p'$ via a linear projection, i.e., $P_i \to P_i' \in \mathbb{R}^{C \times p'}$.

Subsequently, positional encoding is applied to each block to preserve temporal information lost during the segmentation process. The positional embedding of the $i$-th block is denoted as $E_i$, and the output of this step is represented as $X_{PE_i} = P_i' + E_i$.

The encoded features are converted from the frequency domain to the spatial domain through fast Fourier transform (FFT), thereby learning spatial information through global circular convolution. At the same time, adaptive local filters are applied to suppress low-frequency noise. Specifically, given a discrete-time sequence $x[n]$, its frequency-domain representation $X[k]$ is obtained via Fourier transform. Accordingly, for a given positional encoding $X_{PE}$ , its frequency-domain representation is computed as follows:

$$F = FFT(X_{PE}) \in \mathcal{C}^{C \times N'} \tag{1}$$

where $FFT(\cdot)$ denotes the one-dimensional Fast Fourier Transform operation, and $N'$ represents the length of the transformed sequence in the frequency domain. Due to the implementation details of FFT and the characteristics of the input signal, the output length $N'$ may differ from the original signal length. Each channel of the input signal is independently transformed using FFT, resulting in a comprehensive frequency-domain representation $F$, which captures the spectral characteristics of the original signal across all channels.

The high-frequency noise components of the original signal often manifest as rapid fluctuations of the target and exhibit strong randomness. Given the frequency domain representation $F$ obtained by the FFT operation, we first compute its power spectrum to identify the dominant frequency components. The relationship is given by:

$$P = |F|^2 \tag{2}$$

By introducing a learnable threshold $\theta$, low-frequency components can be adaptively suppressed from the power spectrum $P$. This threshold is dynamically adjusted based on the spectral characteristics of the input signal. The filtering process satisfies the following condition:

$$F_{filtered} = F \odot (P > \theta) \tag{3}$$

where $\odot$ denotes the Hadamard (element-wise) product, and $(P > \theta)$ represents the thresholding mask, which retains frequency components with power values above the threshold $\theta$. The adaptive threshold effectively removes irrelevant noise features while preserving critical information. By selecting the frequency threshold in an adaptive manner, the filtering process can be customized for each specific sample, thereby enhancing the overall effectiveness of the model when handling diverse data scenarios.

After adaptive frequency-domain filtering, two sets of learnable filters are introduced: a global filter and a local filter. The global filter is learned from the original frequency-domain representation $F$, while the local filter is derived from the adaptively filtered representation $F_{filtered}$. $\rho_G$ and $\rho_L$ are used to represent the global and local filters, respectively. The filtering process can be formulated as follows:

$$\begin{aligned} F_G &= \rho_G \odot F \\ F_L &= \rho_L \odot F_{filtered} \end{aligned} \tag{4}$$

The final frequency-domain feature representation of the sample signal can be expressed as $F_{fusion} = F_G + F_L$.

Finally, the inverse Fast Fourier Transform (IFFT) is applied to convert the high-dimensional spatial-frequency domain data back to the spatial-temporal domain, yielding the time-domain signal. The resulting time-domain signal is given by:

$$x_{a,t_a}^{'} = IFFT(F_{fusion}) \in R^{C_a \times N_a}. \tag{5}$$

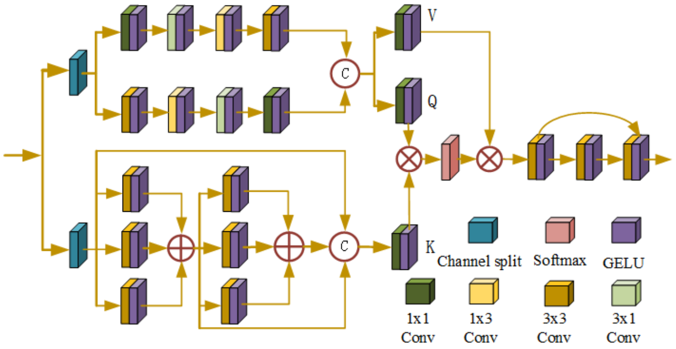

Fig. 3. **Multi-scale visual feature enhancement.**The top-left section corresponds to the dual-channel branch, the bottom-left section represents the triple-channel branch, and the right side is the fusion and enhancement part.

### *B. Multi-scale Visual Feature Enhancement*

In multi-modal audio-video deepfake detection networks, the expressive power of features is critical for improving detection accuracy. To this end, we propose a multi-scale feature enhancement module to augment feature representations by capturing information across different scales, thereby boosting forgery detection performance.

For the visual feature extraction network, preliminary single-frame image features $x_{v,t_v} \in R^{C_v \times H_v \times W_v}$ are first obtained using a pre-trained ResNet-18 backbone, where $C_v$, $H_v$ and $W_v$ denote the number of channels, height, and width of the image features, respectively. As illustrated in Figure 3, the multi-scale feature aggregation network comprises two branches: a dual-channel separation branch and a triple-channel separation branch. After multi-scale feature enhancement, the dual-channel separated features serve as Key and Value, while the triple-channel separated features act as Query. These features are then fused through an attention mechanism.

For the dual-channel split branch, $x_1^{KV} \in R^{C_v/2 \times H_v \times W_v}$ and $x_2^{KV} \in R^{C_v/2 \times H_v \times W_v}$denote the visual features obtained after channel-wise splitting, which can be represented as:

$$x_1^{KV}, x_2^{KV} = Split_2(x_{v,t_v}) \tag{6}$$

The segmented features $x_1^{KV}$ and $x_2^{KV}$ are processed through multi-scale convolutional kernels (as illustrated in the top-left of Figure 3) to extract enhanced features $\tilde{x}_1^{KV}$ and $\tilde{x}_2^{KV}$. These dual-branch enhanced features are then concatenated to yield the final representation:

$$\tilde{x}^{KV} = \text{Concat}\left(\tilde{x}_1^{KV}, \tilde{x}_2^{KV}\right) \tag{7}$$

For the triple-channel split branch, $x_1^Q \in R^{C_v/3 \times H_v \times W_v}$, $x_2^Q \in R^{C_v/3 \times H_v \times W_v}$, and $x_3^Q \in R^{C_v/3 \times H_v \times W_v}$ denote the visual features obtained after channel-wise splitting. Similar to the dual-channel branch, these features can be processed as follows:

$$x_1^Q, x_2^Q, x_3^Q = Split_3(x_{v,t_v}) \tag{8}$$

The segmented features $x_1^Q$, $x_2^Q$ and $x_3^Q$ are processed by a triple-channel convolution layer for feature extraction and then aggregated (as illustrated in the bottom-left of Figure 3) to obtain the feature $x_{123}^Q$. This feature is further processed through another triple-channel convolutional layer with summation to produce the enhanced feature $\tilde{x}_{123}^Q$. Finally, the enhanced features from all three branches are concatenated to form the output representation:

$$\tilde{x}^{Q} = \text{Concat}\left(x_{v,t_v},\ x_{123}^Q,\ \tilde{x}_{123}^Q\right) \tag{9}$$

For feature fusion enhancement, after obtaining the enhanced features from both the dual-channel branch and the triple-channel branch, we apply 1×1 convolutions to generate the Query, Keys, and Value:

$$\begin{aligned} Q &= GELU(Conv_{1\times1}(\tilde{x}^Q)) \\ K &= GELU(Conv_{1\times1}(\tilde{x}^{KV})) \\ V &= GELU(Conv_{1\times1}(\tilde{x}^{KV})) \end{aligned} \tag{10}$$

Then, the aggregated feature can be obtained by applying a standard attention mechanism:

$$\tilde{x}^{QKV} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{11}$$

Finally, the aggregated feature $\tilde{x}^{QKV}$ is enhanced to obtain the enhanced feature $x'_{v,t_v}$ through the feature enhancement part shown on the right side of Figure 3. The feature enhancement part consists of a 3x3 convolution layer, a GELU activation function, and a residual connection.

Multi-scale convolution can effectively capture local details and global structural information in input data. Feature fusion strategies at different scales can fully utilize the complementarity of features at different scales to improve feature discriminability. Through multi-scale feature enhancement, the model aims to improve feature discriminability and mitigate sensitivity to certain perturbations.

### C. Cross-level Multi-modal Fusion Module

In the task of audio-video deepfake detection, effectively integrating feature information from both audio and video modalities is crucial for improving detection accuracy. To this end, we propose a cross-level multi-modal fusion module, which enhances the interaction and complementarity of different modal features through multi-level feature fusion and attention mechanisms. As shown in Figure 4, after obtaining the enhanced audio features $x'_{a,t_a}$ and visual features $x'_{v,t_v}$, unlike previous methods that directly concatenate the features, our approach leverages the coupling between audio and visual features to achieve multi-modal fusion via an attention mechanism. By enhancing the correlation between audio features and visual features, audio information is propagated to different layers of the visual feature representation. Meanwhile, cross-level enhancement is applied to the fused features at various layers, enabling scale-variant feature modeling. The proposed cross-level multi-modal fusion module consists of a multi-modal fusion module and a cross-level enhancement structure.

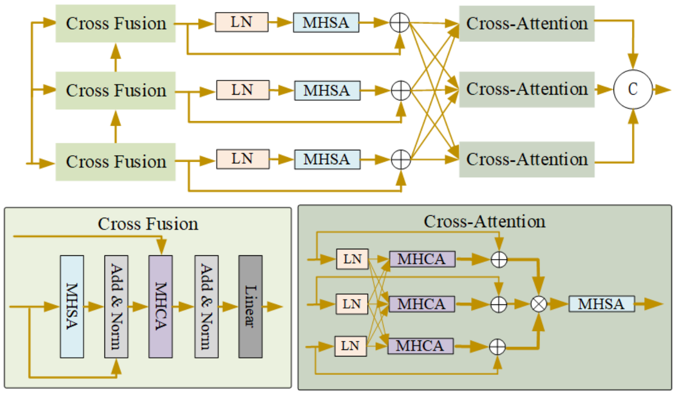

Fig. 4. Cross-level Multi-modal Fusion Module.

The cross-level multi-modal fusion module consists of three similar fusion modules, each sequentially comprising multi-head self-attention (MHSA) and multi-head cross-attention (MHCA) mechanism. MHSA is used to capture long-range dependencies, while MHCA propagates the semantic information embedded in the audio modality into the video modality. In addition, a cross-connected residual structure is added after both MHSA and MHCA. The multi-modal fusion features from the three levels are further enhanced through a residual MHSA, refining the single-stage features. Cross-level cross-attention is then applied for inter-level enhancement. The proposed cross-level multi-modal fusion module is composed of three parts: cross-modal feature fusion, self-attention enhancement, and cross-level enhancement.

**Cross-Modal Feature Fusion.** First, audio and visual features are separately processed through a MHSA mechanism. MHSA captures long-range dependencies within features and enhances their representational capacity. For each modality, MHSA operates in parallel across multiple attention heads, with each head learning a different subspace representation of the features. The features output from MHSA are added to the original features via residual connections, followed by layer normalization. After normalization, the features are fed into a MHCA layer. MHCA computes correlation weights between the audio and visual features to enable cross-modal feature interaction and fusion. Finally, the MHCA-processed features are transformed through a linear layer. By combining MHSA and MHCA, the module can effectively enhance the representational capacity of both audio and visual features, improving their discriminative power. MHCA captures the complementarity between audio and visual features, emphasizing important cross-modal information through weighted aggregation.

**Self-Attention Enhancement.** For the fused features at each level, layer normalization is first applied to stabilize the training process and accelerate convergence. Subsequently, the features are fed into a MHSA layer, which captures long-range dependencies among features and enhances their representational capacity.

**Cross-Level Enhancement.** In a multi-modal learning framework, effectively integrating and utilising information from different levels is crucial for improving task performance. The cross-level cross-attention enhancement module enhances the interactivity and complementarity between features from different levels through a multi-level cross-attention mechanism. Input features are first processed through layer normalization. The normalised features then enter the cross-attention layer. MHCA achieves cross-level interaction and fusion by calculating the correlation weights between features from different levels. The features processed by MHCA are added to the original features via residual connections, followed by further feature fusion. The cross-layer enhanced features are further processed by MHSA, and the final output is the fusion feature $x'_{av,t_{av}}$.

### *D. Learning Objective*

The design of our loss function comprises three components: (1) joint prediction loss for fused audio-visual features; (2) contrastive loss between audio and visual feature outputs; (3) classification loss for individual unimodal (audio/visual) features. These three losses are aggregated via weighted summation to form the final objective function.

We employ a fully connected layer (FC) to project the enhanced audio and visual features into the label space, using a sigmoid function to output classification probabilities. Let $y_a$ and $y_v$ denote the unimodal classification results for the audio and visual modalities, respectively, and $y_{av}$ represent the classification result from the fused multi-modal features. For the $t$-th frame of the video with input features $x'_{a,t_a}$, $x'_{v,t_v}$, and $x'_{av,t_{av}}$, their corresponding predictions are:

$$\begin{aligned} y'_{a,t_a} &= sigmoid(FC(x'_{a,t_a})) \\ y'_{v,t_v} &= sigmoid(FC(x'_{v,t_v})) \\ y'_{av,t_{av}} &= sigmoid(FC(x'_{av,t_{av}})) \end{aligned} \tag{12}$$

The classification loss for the fused features is formulated as a binary cross-entropy (BCE) loss, expressed as:

$$\begin{aligned} L_1 = -\frac{1}{T}\sum_{t=1}^{T}\big(\big(y_{av,t_{av}}\log y'_{av,t_{av}}\big) \\ + (1 - y_{av,t_{av}})\log\big(1 - y'_{av,t_{av}}\big)\big) \end{aligned} \tag{13}$$

A contrastive loss is computed between the audio and visual features to maximize the discrepancy scores for forged samples while minimizing the differences for genuine samples:

$$L_2 = -\frac{1}{T}\sum_{t=1}^{T}\big(y_{av,t_{av}} d_t^2 \; + (1 - y_{av,t_{av}})\max(\lambda - d_t, 0)^2\big) \tag{14}$$

where $d_t = \left\|x'_{v,t_v} - x'_{a,t_a}\right\|$ represents the inter-modal dissonance score, calculated as the Euclidean distance between the visual and audio features, and $\lambda$ is a hyperparameter controlling the loss weight.

The loss of difference between the single modal data labels and the real labels output by the model is used to calculate the loss, which enables the model to learn the single modal data features and improve the differentiation performance of the forged videos. The unimodal loss function is expressed as

$$\begin{aligned} L_3 &= -\frac{1}{T}\sum_{t=1}^{T}\big(y_{a,t_a}\log y'_{a,t_a} + (1 - y_{a,t_a})\log(1 - y'_{a,t_a})\big) \\ L_4 &= -\frac{1}{T}\sum_{t=1}^{T}\big(y_{v,t_v}\log y'_{v,t_v} + (1 - y_{v,t_v})\log(1 - y'_{v,t_v})\big) \end{aligned} \tag{15}$$

The total loss to optimize the proposed audio-video deepfake detection method is:

$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3(L_3 + L_4) \tag{16}$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are the weights for each loss.

## IV. Experiment

### *A. Experimental Setup*

We evaluate our method on the public audio-video deepfake detection dataset FakeAVCeleb [27]. FakeAVCeleb contains 500 real videos and over 20,000 fake videos. Consistent with MRDF [7], a balanced sampling strategy is employed with a 1:1:1:1 ratio across four categories: FakeAudio-FakeVideo (FAFV), FakeAudio-RealVideo (FARV), RealAudio-FakeVideo (RAFV), and RealAudio-RealVideo (RARV). A five-fold cross-validation protocol is adopted.

ResNet-18 is used as the frame-level encoder for both audio and video modalities. Training is performed with the Adam optimizer for 30 epochs, an initial learning rate of 1e-3, and a batch size of 64. Equal weights are assigned to audio classification loss, video classification loss, and fusion classification loss to ensure balanced optimization.

### *B. Overall Performance*

To evaluate the performance of our model, we compare it against ten baseline methods, including AVBYOL [28], Intra-Cross-Modal [29], AD-DFD [9], FACTOR [30], VFD [31], AVOiD-DF [8], DST-Net [32], Ensemble [33], Multimodal-trace [34], and MRDF [7]. To ensure fairness, all models are studied from a multi-modal perspective. The evaluation metrics used are accuracy (ACC) and area under the curve (AUC). Accuracy reflects the model's classification correctness

TABLE I
PERFORMANCES COMPARISON OF AUDIO-VIDEO DEEPFAKE DETECTION WITH OTHER METHODS ON FAKEAVCELEB.

| **Method** | **ACC(%)** | **AUC(%)** |
|---|---|---|
| VFD [31] | 81.52 | 86.11 |
| Multimodaltrace [34] | 92.90 | - |
| Ensemble [33] | 89.00 | - |
| DST-Net [32] | 92.59 | - |
| AVOiD-DF [8] | 83.70 | 89.20 |
| AVBYOL [28] | 92.16 | 90.41 |
| Intra-Cross-Modal [29] | 93.52 | 91.55 |
| AD-DFD [9] | 91.89 | 90.11 |
| FACTOR [30] | 90.97 | 89.12 |
| MRDF [7] | 94.05 | 92.43 |
| AURORA(ours) | **94.32** | **93.66** |

between forged and real samples; the closer ACC is to 1, the better the classification performance. Similarly, a higher AUC value, approaching 1, indicates better discrimination between positive and negative samples across all possible threshold settings.

As shown in Table I on the FakeAVCeleb benchmark dataset, the proposed AURORA method significantly outperforms existing representative approaches in audio-video deepfake detection tasks. Specifically, AURORA achieves an accuracy (ACC) of 94.32% and an area under the curve (AUC) of 93.66%, improving over the second-best method MRDF (ACC 94.05%, AUC 92.43%) by 0.27% and 1.23%, respectively. Compared to earlier models based on single modality or simple fusion strategies such as VFD and AVOiD-DF, AURORA leads by more than 7.55% in terms of AUC, validating the effectiveness of the cross-modal fine-grained collaborative mechanism. Furthermore, AURORA surpasses the latest multi-modal methods including Multimodaltrace, Intra-Cross-Modal, and MRDF on both ACC and AUC metrics, further demonstrating the significant advantages of the proposed adaptive time-frequency feature aggregation and cross-level multi-modal enhancement in suppressing inter-modal redundancy and enhancing the perception of forgery traces.

### *C. Ablation Study*

Table II presents the ablation study results on the FakeAVCeleb dataset, evaluating the incremental contributions of the core components of AURORA. Using MRDF (ACC 94.05%, AUC 92.43%) as the baseline, AAFE, MVFE, and multi-modal fusion modules are introduced sequentially to observe their effects on detection performance.

With AAFE integrated alone, ACC increases from 94.05% to 94.09%, and AUC improves from 92.43% to 92.52%, validating the effectiveness of introducing adaptive time-frequency attention in the audio branch for capturing transient distortions in forged speech. When using MVFE alone, ACC reaches 94.08% and AUC rises to 92.59%, significantly outperforming AAFE alone. This indicates that the visual branch captures facial forgery artifacts at different granularities through multi-branch channel-separated aggregation.

TABLE II
ABLATION STUDY OF THE PROPOSED METHOD ON FAKEAVCELEB.

| **Baseline** | **AAFE** | **MVFE** | **Multi-modal Fusion** | **ACC(%)** | AUC(%) |
|---|---|---|---|---|---|
| ✓ | - | - | - | 94.05 | 92.43 |
| ✓ | ✓ | - | - | 94.09 | 92.52 |
| ✓ | - | ✓ | - | 94.08 | 92.59 |
| ✓ | - | - | ✓ | 94.17 | 92.98 |
| ✓ | ✓ | ✓ | - | 94.28 | 93.21 |
| ✓ | ✓ | ✓ | ✓ | **94.32** | **93.66** |

When only the cross-level cross-modal fusion module is introduced, ACC and AUC jump to 94.17% and 92.98%, respectively, substantially surpassing single-modality enhancement strategies. This highlights the critical role of cross-modal information interaction in suppressing modal redundancy and enhancing consistency of forgery traces.

When all three modules collaborate, the model achieves optimal performance, with ACC reaching 94.32% and AUC reaching 93.66%, corresponding to relative improvements of 0.29% and 1.23% over the baseline. This fully demonstrates the complementarity of the three modules along the "audio–visual–fusion" pipeline: AAFE and MVFE extract more discriminative micro- and macro-level forgery features within their respective modalities, while multi-modal fusion achieves cross-level alignment and complementarity in the high-level semantic space, ultimately enabling robust detection of deepfake videos.

## V. LIMITATIONS

Although AURORA has achieved significant performance improvements in audio-video deepfake detection tasks, it still exhibits certain limitations. The model architecture is relatively complex and incurs substantial computational overhead. Future work could explore more lightweight network designs to meet the requirements of real-time detection, particularly for deployment on resource-constrained devices (e.g., mobile platforms).

Additionally, the current model has been validated primarily on the FakeAVCeleb dataset. Although this dataset contains a wide range of forgery types, in the real world, forgery techniques are constantly evolving and the diversity of forgery samples far exceeds the current dataset. In the future, the size of the dataset can be further expanded to cover more types of forgery samples to improve the generalisation ability of the model.

## VI. CONCLUSION

In this paper, we address a key challenge in audio–visual deepfake detection—namely, the fine-grained capture and consistent fusion of cross-modal forgery traces—by proposing the AURORA and conducting a systematic evaluation on the public FakeAVCeleb benchmark. In terms of overall performance, AURORA achieves an ACC of 94.32% and an AUC of 93.66%, surpassing the next-best approach by 0.27% and 1.23%, respectively, demonstrating robustness in high-fidelity forgery scenarios. The experimental results clearly indicate that, in audio–video deepfake detection, relying solely on a single modality or shallow cross-modal fusion is insufficient to counter increasingly sophisticated forgery techniques. By combining intra-modal enhancement with cross-modal alignment in a hierarchical design, the proposed approach suppresses redundant noise while maximizing the cross-modal consistency of forgery traces, offering a promising technical pathway for next-generation anti-forgery systems.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62472083, 62432008) and the AI-Enhanced Research Program of Shanghai Municipal Education Commission (Grant No. SMEC-AI-DHUZ-01).

## References

[1] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] J. He, L. Zhang, Y. Zhu, and L. Dong, "Mcdm-dti: Accelerating diffusion tensor imaging based on multi-conditional denoising diffusion probability model," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pp. 186–191, IEEE, 2024.

[4] M. Z. Uddin, M. A. Shahriar, M. N. Mahamood, F. Alnajjar, M. I. Pramanik, and M. A. R. Ahad, "Deep learning with image-based autism spectrum disorder analysis: A systematic review," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107185, 2024.

[5] M. Javed, Z. Zhang, F. H. Dahri, and A. A. Laghari, "Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach," *Electronics*, vol. 13, no. 15, p. 2947, 2024.

[6] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, pp. 7–15, 2021.

[7] H. Zou, M. Shen, Y. Hu, C. Chen, E. S. Chng, and D. Rajan, "Cross-modality and within-modality regularization for audio-visual deepfake detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4900–4904, IEEE, 2024.

[8] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

[9] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14800–14809, 2021.

[10] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25494–25513, 2022.

[11] M. S. Rana, B. Murali, and A. H. Sung, "Deepfake detection using machine learning algorithms," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 458–463, IEEE, 2021.

[12] H. Wen, S. Chang, L. Zhou, W. Liu, and H. Zhu, "Opticloak: Blinding vision-based autonomous driving systems through adversarial optical projection," *IEEE Internet of Things Journal*, vol. 11, no. 17, pp. 28931–28944, 2024.

[13] A. K. Jha, A. K. Yadav, A. K. Dubey, A. Kumar, and A. Sharma, "Deep learning based deepfake video detection system," in *2025 3rd International Conference on Disruptive Technologies (ICDT)*, pp. 408–412, IEEE, 2025.

[14] J. Chen, W. Lin, and J. Xu, "Deepfake detection using graph representation with multi-dimensional features," in *2023 IEEE Smart World Congress (SWC)*, pp. 717–722, IEEE, 2023.

[15] H. Wen, S. Yan, S. Chang, J. Xu, H. Zhu, Y. Zhang, and B. Li, "Depth-cloak: Projecting optical camouflage patches for erroneous monocular depth estimation of vehicles," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2739–2747, 2024.

[16] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[17] S. Chang, Y. Qi, H. Zhu, J. Zhao, and X. Shen, "Footprint: Detecting sybil attacks in urban vehicular networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 6, pp. 1103–1114, 2011.

[18] S. Muppalla, S. Jia, and S. Lyu, "Integrating audio-visual features for multimodal deepfake detection," in *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–5, IEEE, 2023.

[19] Y. Yang, L. Yuan, J. Zhao, and W. Gong, "Content-agnostic backscatter from thin air," in *Proceedings of the 20th annual international conference on mobile systems, applications and services*, pp. 343–356, 2022.

[20] M. Javed, F. H. Dahri, A. A. Laghari, J. A. Bhutto, S. H. Bhutto, and N. A. Dahri, "Mmdf-net: A multimodal framework for real-time deepfake detection using visual and audio features," in *2025 IEEE 2nd International Conference on Electronics, Communications and Intelligent Science (ECIS)*, pp. 1–5, IEEE, 2025.

[21] Q. Wang, S. Chen, J. Zhao, and W. Gong, "Rapidrider: Efficient wifi backscatter with uncontrolled ambient signals," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2021.

[22] S. Yan, H. Wen, S. Chang, H. Zhu, and L. Zhou, "Fooling 3d face recognition with one single 2d image," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4043–4052, 2024.

[23] Z. Jin, L. Lang, and B. Leng, "Wave-spectrogram cross-modal aggregation for audio deepfake detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[24] W. Gong, L. Yuan, Q. Wang, and J. Zhao, "Multiprotocol backscatter for personal iot sensors," in *Proceedings of the 16th international conference on emerging networking experiments and technologies*, pp. 261–273, 2020.

[25] C. Shirley, B. J. Jingle, M. Abisha, R. Venkatesan, Y. R. RV, E. Elango, *et al.*, "Deepfake detection using multi-modal fusion combined with attention mechanism," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, pp. 1194–1199, IEEE, 2024.

[26] L. Yuan, C. Xiong, S. Chen, and W. Gong, "Embracing self-powered wireless wearables for smart healthcare," in *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–7, IEEE, 2021.

[27] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[28] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14950–14962, 2022.

[29] M. Tian, M. Khayatkhoei, J. Mathai, and W. AbdAlmageed, "Unsupervised multimodal deepfake detection using intra-and cross-modal inconsistencies," *arXiv preprint arXiv:2311.17088*, 2023.

[30] T. Reiss, B. Cavia, and Y. Hoshen, "Detecting deepfakes without seeing any," *arXiv preprint arXiv:2311.01458*, 2023.

[31] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–22, 2023.

[32] H. Ilyas, A. Javed, and K. M. Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," *Applied Soft Computing*, vol. 136, p. 110124, 2023.

[33] A. Hashmi, S. A. Shahzad, W. Ahmad, C. W. Lin, Y. Tsao, and H.-M. Wang, "Multimodal forgery detection using ensemble learning," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1524–1532, IEEE, 2022.

[34] M. A. Raza and K. M. Malik, "Multimodaltrace: Deepfake detection using audiovisual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 993–1000, 2023.