AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection

Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, *Fellow, IEEE*, Yu Tsao, *Senior Member, IEEE*, and Hsin-Min Wang, *Senior Member, IEEE*

***Abstract*—The recent proliferation of hyper-realistic deepfake videos has drawn attention to the threat of audio and visual forgeries. Most previous studies on detecting artificial intelligence-generated fake videos only utilize visual modality or audio modality. While some methods exploit audio and visual modalities to detect forged videos, they have not been comprehensively evaluated on multimodal datasets of deepfake videos involving acoustic and visual manipulations, and are mostly based on convolutional neural networks with low detection accuracy. Considering that human cognition instinctively integrates multisensory information including audio and visual cues to perceive and interpret content and the success of transformer in various fields, this study introduces the audio-visual transformer-based ensemble network (AVTENet). This innovative framework tackles the complexities of deepfake technology by integrating both acoustic and visual manipulations to enhance the accuracy of video forgery detection. Specifically, the proposed model integrates several purely transformer-based variants that capture video, audio, and audio-visual salient cues to reach a consensus in prediction. For evaluation, we use the recently released benchmark multimodal audio-video FakeAVCeleb dataset. For a detailed analysis, we evaluate AVTENet, its variants, and several existing methods on multiple test sets of the FakeAVCeleb dataset. Experimental results show that the proposed model outperforms all existing methods and achieves state-of-the-art performance on Testset-I and Testset-II of the FakeAVCeleb dataset. We also compare AVTENet against humans in detecting video forgery. The results show that AVTENet significantly outperforms humans.**

***Index Terms*—Deepfake detection, audio-visual, AVTENet**

I. Introduction

WIDESPREAD misinformation or disinformation [1] on online social media has fueled the spread of unverified rumors, often escalating into social disputes. Deepfake technology further exacerbates this issue. Deepfake refers to forged media content in which the face and/or voice of the source person are manipulated with the face and/or voice of a target person using deep learning algorithms. Deepfake detection is essential for applications such as digital forensic analysis, content moderation, and protecting public trust in media integrity.

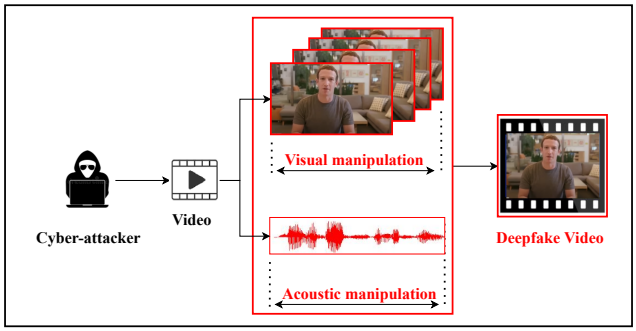


Fig. 1. Example of audio-visual deepfakes. Cyber-attackers aim to create convincing counterfeit videos by manipulating audio and visual streams. Sharing such malicious content can have harmful consequences. Therefore, our proposed method exploits manipulations in both audio and visual streams to effectively detect forged videos.

Modern deep learning technology has significantly enhanced the generation or manipulation of media content through various advanced tools and techniques. Generative adversarial networks (GANs) [2] and variational autoencoders (VAEs) [3] can create highly realistic synthetic media content that is indistinguishable from real content. Face swapping, face reenactment, face editing, and face synthesis are various categories of deepfake techniques that help produce convincing and well-crafted facial deepfake media [4]. Initially, deepfake content involved unimodal manipulation of audio or video. But recently, multimodal manipulation has been used to produce more realistic deepfake videos. The viral Instagram video of Mark Zuckerberg (Meta's CEO), as shown in Fig. 1, is the ultimate example of audio-visual deepfakes that combine acoustic and visual forgeries.

Detecting such convincing deepfake videos is crucial and challenging. To combat the trend of deepfakes, several deep learning-based solutions have been proposed to detect forgeries in videos. Existing deepfake detection methods only focus on unimodal detection of video deepfakes, lacking evaluation on multimodal deepfake datasets. Unimodal methods limit the performance of detectors to certain scenarios; for instance, an audio-only detector (Fig. 2(a)) cannot detect visual manipulation, while a detector that relies solely on the visual modality (Fig. 2(b)) can easily be fooled by acoustic manipulation.

The human brain inherently combines auditory and visual cues to comprehend the world. This multisensory cognitive

Ammarah Hashmi is with the Social Networks and Human-Centered Computing Program, Taiwan International Graduate Program, Academia Sinica, Taipei 11529, Taiwan, and also with the Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu 30013, Taiwan. (e-mail: hashmiammarah0@gmail.com).

Sahibzada Adil Shahzad is with the Social Networks and Human-Centered Computing Program, Taiwan International Graduate Program, Academia Sinica, Taipei 11529, Taiwan, and also with the Department of Computer Science, National Chengchi University, Taipei 11605, Taiwan. (e-mail: adilshah275@iis.sinica.edu.tw).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan. (e-mail: cwlin@ee.nthu.edu.tw)

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan. (e-mail: yu.tsao@citi.sinica.edu.tw).

Hsin-Min Wang is with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan. (Corresponding Author, e-mail: whm@iis.sinica.edu.tw)

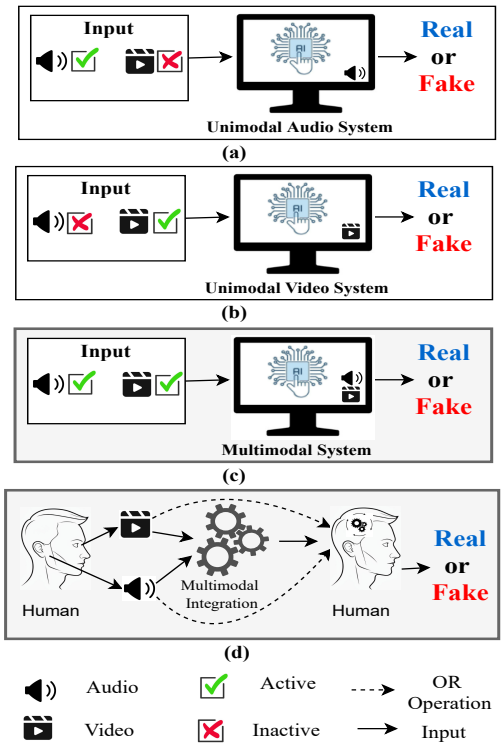


Fig. 2. Comparison of artificial systems and humans in detecting video forgery.

ability [5], [6] or multimodal data [7] brings greater richness and representation and helps identify inconsistencies or anomalies in deepfake media, as shown in Fig. 2(d). Therefore, by drawing inspiration from human cognitive skills and attention mechanisms, our approach bridges the gap between natural and artificial systems, leveraging auditory and visual sensory information [7] and transformer networks [8] to detect deepfake videos (Fig. 2(c)), and is evaluated on a multimodal dataset to take a step toward applications in real-world scenarios.

Recently, transformers [8] have emerged as a promising alternative and marked a dominant shift from convolutional neural networks (CNNs) in computer vision tasks due to their ability to model long-range dependencies and capture global context. The sequential processing nature of CNN limits the ability of the network to capture long-range dependencies; however, the attention mechanism in the transformer allows it to model long-range global context and relationships. Existing methods mainly use CNN-based solutions to deal with deepfake detection [9]–[16]. These CNN-based methods process input data sequentially, while transformers can process it efficiently by processing the entire input sequence in parallel. Furthermore, in natural language processing (NLP) and computer vision, many transformer-based foundation models trained with large amounts of data have proven useful for downstream application tasks, such as the video vision transformer (ViViT) [17], the audio spectrogram transformer (AST) [18], and the audio-visual transformer (AV-HuBERT) [19]. Due to the effectiveness of transformers in various fields [20]–[22], we intend to apply them to the video forgery detection task by leveraging acoustic and visual features. We believe that multiple modalities provide complementary information and stable inferences [7], [23], [24]. Therefore, to fill the above gap, we propose a novel **A**udio-**V**isual **T**ransformer-based **E**nsemble **Net**work, **AVTENet**, that utilizes both acoustic and visual features for effective video forgery detection.

Our innovation lies in the application and integration of cutting-edge technologies that were not originally proposed for audio-visual deepfake detection. The main contributions of our work are summarized as follows:

- We propose an audio-visual transformer-based ensemble network (AVTENet), and demonstrate that its audio-only, video-only, and audio-visual networks respectively based on cutting-edge pre-trained transformer-based foundation models can be effectively combined to improve robustness and detection accuracy. The proposed approach can address scenarios where only a single modality or both modalities in audio vision are manipulated.
- To evaluate the strengths and weaknesses of various ensemble approaches, we present four variants of AVTENet and show that the feature fusion approach performs better than majority voting, score fusion, and average score fusion.
- We experimentally confirm the effectiveness of AVTENet and empirically demonstrate that our transformer-based ensemble model outperforms the previous CNN-based ensemble network [9].
- To evaluate the proposed method, we use the FakeAVCeleb dataset [25], which is a multimodal audio-visual deepfake dataset. We conduct a comprehensive experimental comparison between the proposed method and existing unimodal and multimodal methods. The results confirm that our proposed method achieves the highest performance on the test sets of the FakeAVCeleb dataset, outperforming all compared methods.
- We compare AVTENet's detection capabilities with human detection capabilities.

The remainder of this paper is organized as follows. Sec. II reviews related work. Sec. III illustrates the proposed methodology. Sec. IV outlines the experimental setup and reports the results. Sec. V discusses the limitations of this study. Finally, Sec VI provides conclusions and future work.

II. Related Work

A. Video Deepfake Generation

Deepfake technology has become more sophisticated as it uses AI techniques to generate, manipulate, or alter video content, thereby adding significant threats to society, legislation, individuals, community consensus, journalism, and cyber-security. Face2Face [26], Neural Talking-Heads [27], First Order Motion Model for Image Animation [28], and Deep Video Portraits [29] are some of the deepfake generation techniques. Likewise, there exist different types of deepfake videos, including lip-syncing, face-swapping, and voice manipulation. GANs [2], [30] have become a popular deep-learning method for generating realistic images or videos. For instance, in [27] and [31] GAN-based video deepfake generation methods were proposed. The method proposed in [31] generates a dance video of a person from a single image. First,

a sequence of human poses is generated and rendered in the video. To generate realistic animated dance videos, this method combines pose estimation, key-point detection, and image synthesis. Similarly, the few-shot learning method proposed in [27] uses a few reference images to generate realistic videos of talking heads. Their method creates a realistic video sequence of a target person speaking by mapping several reference images. Another GAN-based method for image-to-image translation was proposed in [32]. To ensure high perceptual qualities of the generated images, the model was trained with a perceptual loss function. The authors evaluated their method on several image generation tasks, such as image colorization, painting, and super-resolution.

B. Video Deepfake Detection

1) Unimodal Deepfake Detection: Previous work has mainly focused on a single modality to detect forgeries in videos since synthetic videos generally have less manipulation in the audio track (Fig. 2(b)). Some researchers have worked on detecting forgeries in videos by exploiting visual artifacts such as inconsistent facial expressions and movements [33], [34], irregular head movements or angles [33], [35], mismatched shadows or lightening [36], distorted or blurred edges around the face or body [37], unnatural eye blinking or movement [38], unnatural background or scenery [39]. The authors of [40] argued that several facial editing techniques exhibit visual artifacts in videos that can be tracked and used to detect video forgeries. Considering this rationale, they demonstrated that even simple visual cues can reveal deepfakes. The authors of [38] used eye blinking, a physiological signal that is not well presented in synthetic videos, to detect fake face videos. Other researchers have exploited emotional inconsistency [33], synchronization between speech and lip movements [10], or motion analysis [41]. A biometrics-based forensics technique was proposed in [42] to detect faceswap deepfakes by exploiting appearance and behavior, especially facial expressions and head movements. MesoNet and MesoInception are two popular models proposed in [11]. Both are compact CNN-based models with a few layers that exploit mesoscopic properties to detect facial tampering in videos. MesoNet analyzes the residual high-frequency content of video frames. XceptionNet [12] is built on depth-wise separable convolutional layers. The CNN-based method proposed in [13] employs composite coefficients that effectively scale all dimensions of depth, width, and resolution uniformly.

A transformer network based on an incremental learning strategy is proposed in [43], which utilizes face images and their UV texture generated by a 3D face construction method using a single-face input image to detect synthetic videos. Later, the same authors proposed a hybrid transformer network to detect forged videos [44]. The transformer network is trained end-to-end using XceptionNet and EfficientNet-B4 as feature extractors. Obscene video detection [45] can be considered related to the task of fake video detection, where most obscene videos of celebrities are fake.

2) Multimodal Deepfake Detection: Considering that the auditory context is crucial in visual object perception [46], few recent attempts have addressed the challenge of detecting multimodal deepfake videos using multiple modalities (Fig. 2(c)) such as video, audio, and text. The CNN-based method proposed in [14] utilizes audio and video streams to extract features and then combines these features to better detect deepfake videos. A learning-based approach, inspired by Siamese networks and triplet loss, is proposed in [47]. This approach enhances information for learning by using the acoustic and visual features of the same video. The authors extracted affective cues and used them to compare the emotions perceived in the two modes to discern whether the video was genuine or manipulated. The authors of [48], [49], and [50] combined audio and visual modalities to develop a video deepfake detector that can handle audio-only, video-only, and audio-visual manipulations. In [48], a promising approach was proposed to exploit the intrinsic synchronization between audio and visual modalities to potentially detect increasingly prevalent deepfakes. In [49], an audio-visual person-of-interest deepfake detector that exploits specific biometrics of individual was proposed to handle various manipulation methods used in synthetic media generation. The high generalizability of the detector is due to its independence from any manipulation techniques and training on videos of real talking faces. Similarly, for audio-visual forgery detection, the authors of [50] incorporate audio loss, video loss and audio-visual loss to capture the inconsistency of multimodality and artifacts from individual modalities. Additionally, ensemble approaches were used in [9] and [51] to improve the accuracy of deepfake detection. The ensemble method in [9] combines CNN-based audio-only, video-only, and audio-visual networks to detect multimodal deepfakes. In [51], incorporating supplementary textual features was demonstrated to be effective in handling multimodal fake content detection. The authors built an ensemble network combining unimodal and crossmodal classifiers to distinguish forged from genuine video clips. Furthermore, the authors of [15] argued that the manipulation of audio or visual flow in video can lead to a lack of harmony between them, such as speech-lip non-synchrony and unnatural facial movements. Therefore, they utilized modality dissonance scores to measure the dissimilarity between two modalities to identify genuine or spoofed videos. Another deep learning method that exploits the audio and visual modalities and considers the importance of speech-lip synchronization for the task of video forgery detection was proposed in [10]. Their audio-visual deepfake detector checks the synchrony between the lip sequence extracted from the video and the synthetic lip sequence generated from the audio track. Similarly, the authors of [52] adopted a self-supervised transformer-based approach that utilizes contrastive learning. Their method allows paired acoustic and visual streams to learn mouth motion representations propelling paired representation to be closer and unpaired to stay farther to determine audio-visual forgery. Another self-supervised learning approach in [53] exploits an audio-visual temporal synchronization by evaluating consistency between the acoustic stream and faces in a video clip to determine forgery. In addition to the above studies, there are also multimodal methods for other tasks worth referring to. For instance, [54] proposed a two-stage

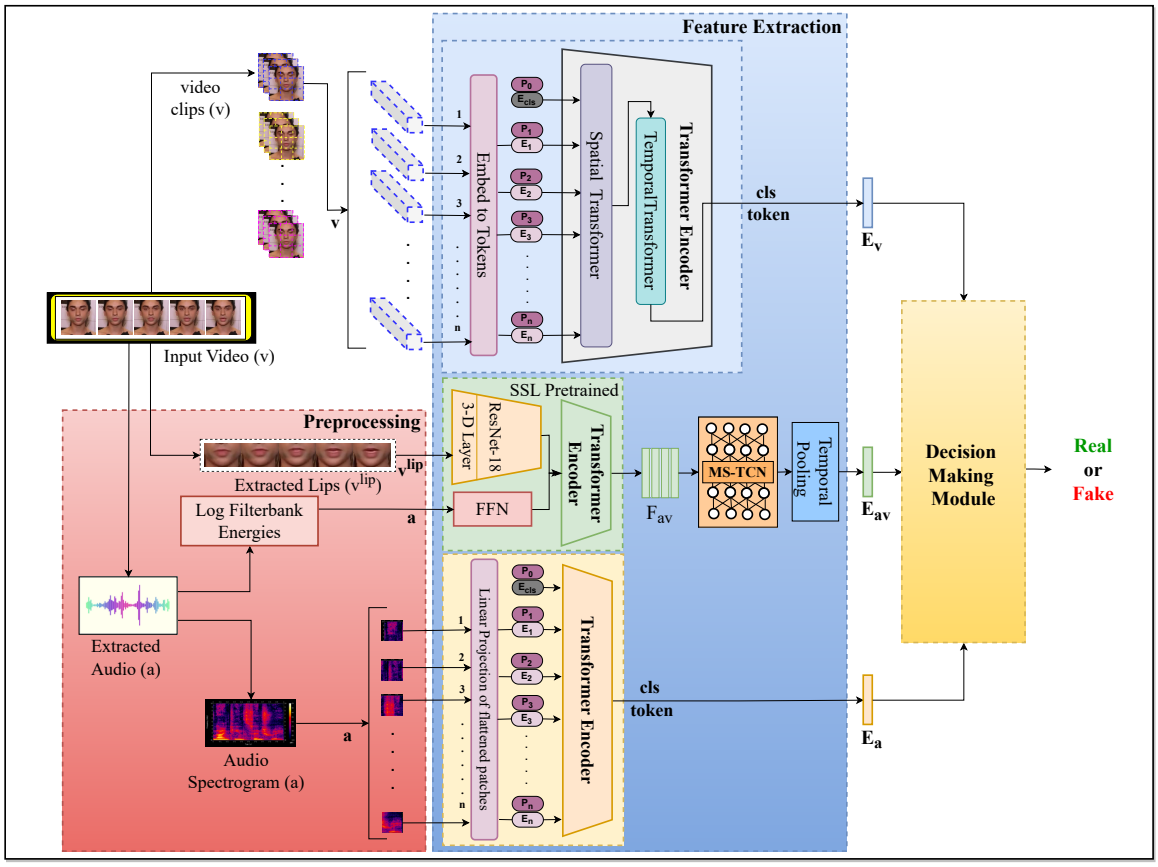


Fig. 3. Overview of the proposed AVTENet model for detecting video forgeries by simultaneously utilizing acoustic and visual cues. It consists of a video-only network (VN), an audio-only network (AN), and an audio-visual network (AVN), each of which is an independent transformer-based classifier, and a decision-making module (DM) that produces final predictions based on the three classifiers through various ensemble strategies.

cascade framework based on multimodal data fusion for face anti-spoofing. The authors of [55] proposed a deep CNN–called LieNet to detect deception by combining contact and non-contact signal modalities. The authors of [56] proposed a multimodal feature-fusion-based model called MFFFLD for fingerprint liveness detection. [57] compared the recognition performance and robustness of different multimodal emotion recognition models. Although multimodal approaches have been studied in video forgery detection; more exploration is needed.

C. Multimodal Deepfake Datasets

To the best of our knowledge, DFDC [58] and FakeAVCeleb [25] are the only multimodal datasets that include both acoustic and visual manipulations in videos.

The DFDC dataset [58] is larger than the FakeAVCeleb dataset, containing 100,000 video clips collected from 3,426 paid subjects. DFDC provides a benchmark for evaluating the performance of deepfake detection models and serves as a crucial resource for advancing research in deepfake detection. It has been released as part of a competition to promote the development of deepfake detection methods. The deepfake videos in this dataset were generated using various deepfake techniques, such as GAN-based methods and nonlearning methods. It comprises real and fake videos and considers both acoustic and visual manipulations.

The FakeAVCeleb dataset [25] is a multimodal audio-visual dataset released after DFDC. The dataset was carefully generated through multiple manipulation methods, considering the balance among ethnic backgrounds, gender, and age groups. The real videos of 500 celebrities originating from the Voxceleb2 dataset [59] collected from YouTube form the base set. This base set was further used to generate 19,500 deepfake videos through manipulation techniques, including Faceswap [60], FSGAN [61], Wav2Lip [62], and RTVC (real-time-voice-cloning) [63].

III. Proposed Method

The natural auditory context provides distinctive, independent, and diagnostic information about the visual world, which directly impacts the perceptual experience of visual objects [6] [46]. To combat the emerging threat of video deepfakes, inspired from human cognition of multisensory information, we leverage both acoustic and visual information to detect forgeries in videos. We use the transformer-based model because its self-attention mechanism can catch inconsistencies or manipulations to identify video forgeries. Furthermore, we employ ensemble learning to exploit the fusion of complementary information, modality-specific patterns, and multimodal features to improve the robustness of the detection system [7]. Accordingly, we propose an AVTENet that fully exploits acoustic and visual information to detect video forgeries by leveraging modality-specific cues and complementary insights

of joint audio and visual information. Fig. 3 shows an overview of our proposed AVTENet model, which consists of three key networks, namely a video-only network (VN), an audio-only network (AN), and an audio-visual network (AVN), as well as a decision-making module (DM). The three key networks are different-modality transformer-based networks integrated with pre-trained models through self-supervised learning (SSL) or supervised learning (SL). Subsequently, DM integrates the outputs of these three classifiers according to different fusion strategies. Given a test video x, the decision is made according to

$$\mathrm{AVTENet}(x) = \mathrm{DM}(C_v(x), C_a(x), C_{av}(x)), \tag{1}$$

where C_v, C_a, and C_{av} denote the video-only, audio-only, and audio-visual classifiers, respectively, and DM denotes the function of decision making.

The AVTENet architecture is designed to effectively capture and utilize complementary information from both modalities. Three networks are used in our ensemble approach to leverage the strength of each network. Individual unimodal networks (AN and VN) specialize in processing information from specific modalities. AN specializes in analyzing audio data to extract valuable acoustic features that help detect acoustic manipulation, while VN focuses on processing video data to extract visual features that can reveal visual forgeries. Combining them into AVN allows for a more comprehensive analysis of the input and can capture complex relationships and dependencies between the two modalities. These multimodal features provide more reliable detection results and enhance robustness.

A. Video-only Network (VN)

The VN module extracts relevant spatiotemporal features from video frames through its self-attention mechanism to capture spatiotemporal patterns and long-range dependencies. It then outputs a vector representation that captures essential visual information.

Building on the remarkable progress achieved by the video vision transformer ViViT [17] in video classification tasks, we seamlessly integrate it into the visual backbone of AVTENet. As shown in Fig. 3, we use a factorized encoder model that consists of two separate transformer encoders. The first is a spatial encoder, which delves into the interaction among tokens extracted from the same temporal index. The second is a temporal encoder that models interactions among tokens from different temporal indices. We first divide the input video into small segments, a series of patches is extracted from each segment as a tubelet and passed through the spatial encoder to produce an encoding vector representing the segment. The output encoding vectors serve as input to the temporal transformer along with an additional classification (cls) token to extract the final representation of the entire video.

For VN training, a dataset $D^v = \{v_i, y_i\}_{i=1}^n$ is extracted from the training set of the FakeAVCeleb dataset, where v_i denotes the video stream of the i-th training sample x_i, and y_i is the label for v_i (0 for fake and 1 for real). As shown in Fig. 3, the ViViT model extracts the single-vector representation E_v (the cls token of the temporal transformer) of each training video clip. A linear layer is used as the classifier. The ViViT model is pre-trained on the Kinetics dataset [64]. During the VN training process, the linear classification layer is trained, and the pre-trained ViViT model is fine-tuned. The model is trained with the binary cross-entropy loss defined as

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i), \tag{2}$$

where n is the total number of training samples, and $\hat{y}_i$ denotes the predicted score.

In the inference phase of VN, given a test sample x, the prediction is conducted by

$$C_v(x) = \mathrm{VN}(x_v), \tag{3}$$

where x_v is the video stream (i.e., the image sequence) of x. In the inference phase of our ensemble model, the DM module utilizes either the single-vector representation or the predicted score corresponding to the "fake" of the video-only network.

B. Audio-only Network (AN)

The AN module takes as input an audio spectrogram, where time and frequency elements help learn and capture acoustic patterns, temporal dynamics, and other audio-specific characteristics, and returns a high-dimensional feature representation.

Transformers perform incredibly well in audio processing as their self-attention mechanisms allow for capturing long-range dependencies in audio. In this study, we use the audio spectrogram transformer (AST) [18] as the audio backbone of AVTENet. As shown in Fig. 3, the AN module takes as input a spectrogram that is further divided into a sequence of 16x16 patches with an overlap of six in both the time and frequency dimensions. A linear projection layer transforms each patch into a 1D embedding. Since the patch sequence lacks input order information, trainable positional embeddings are incorporated into each patch embedding to enable capturing spatiotemporal structure from a 2D audio spectrogram. Additionally, a classification (cls) token is added to the sequence. The output cls embedding of the transformer encoder encapsulates information about characteristics that can be used to determine challenging and deceptive alterations in the audio.

For training AN, a dataset $D^a = \{a_i, y_i\}_{i=1}^n$ is extracted from the training set of the FakeAVCeleb dataset, where a_i denotes the audio track of the i-th training sample x_i, and y_i is the label for a_i (0 for fake and 1 for real). As shown in Fig. 3, the AST model is used to extract the single vector representation E_a (the cls token) of the mel-spectrogram of each training audio clip. A linear layer is used as the classifier. The AST model is pre-trained on the Audioset [18]. During the training of AN, the linear classification layer is trained, and the pre-trained AST model is fine-tuned. As with the video-only network, the cross-entropy loss is used to train the audio-only network.

In the inference phase of AN, given a test sample x, the prediction is conducted by

$$C_a(x) = \mathrm{AN}(x_a), \tag{4}$$

where x_a is the audio track of x. As with the video-only network, in the inference phase of our ensemble model, either the single-vector representation or the predicted score corresponding to the "fake" of the audio-only network is used by the DM module.

C. Audio-Visual Network (AVN)

The AVN module exploits both visual and acoustic information by jointly learning paired audio and visual inputs that provide complementary information and capture meaningful representations between these distinct modalities to expose fabrication in videos.

We use the self-supervised learning framework AV-HuBERT [19] as the audio-visual backbone of AVTENet. As shown in Fig. 3, the AVN module simultaneously uses the log filterbank energies and the video frame stack of the lip region extracted from the video as inputs to the lightweight modality-specific encoders. The frames of the lip region are passed through a ResNet-based feature extractor to extract relevant visual features. Simultaneously, the log filterbank energy features are passed through the feed-forward network (FFN) to extract the acoustic features. The visual and acoustic features are then fused and fed to a shared transformer encoder to extract jointly contextualized audio-visual representations that encapsulate the correlation between acoustic and visual modalities. These audio-visual representations are further passed through a multiscale temporal convolutional network (MS-TCN) and a temporal pooling layer to produce a single-vector representation. Fig. 4 shows the structural diagram of MS-TCN. The details can be found in [65] and [66].

For training AVN, a dataset $D^{av} = \{a_i, v_i^{lip}, y_i\}_{i=1}^{n}$ is extracted from the training set of the FakeAVCeleb dataset, where a_i and v_i^{lip}, respectively, denote the audio track and the lip image sequence of the i-th training sample x_i, and y_i is the label for x_i (0 for fake and 1 for real). As shown in Fig. 3, the AV-HuBERT model, MS-TCN network, and temporal pooling layer are used to extract the vector representation E_{av} of the log filterbank energy features of the audio track and the lip-image sequence of each training video clip. A linear layer is used as the classifier. The AV-HuBERT model is pre-trained

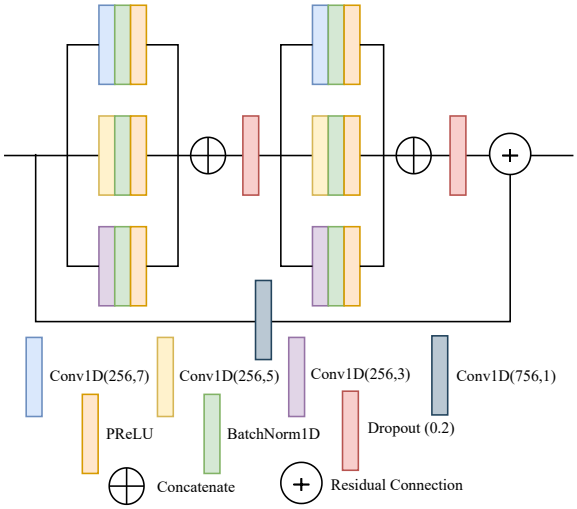


Fig. 4. Architecture of MS-TCN block.

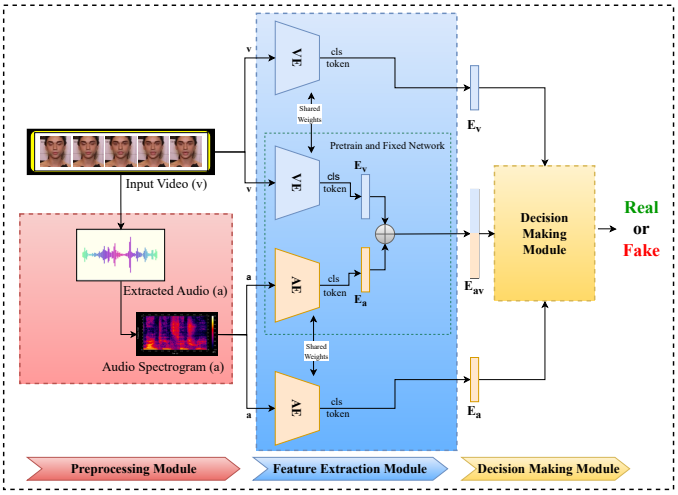


Fig. 5. Overview of another AVTENet model that consists of AST-based AN, ViViT-based VN, and AST&ViViT-based AVN.

on the LRS3 dataset [67]. During the training process of AVN, both the MS-TCN with a linear classification layer and the pre-trained AV-HuBERT model are trained, with the latter being fine-tuned.

In the inference phase of AVN, given a test sample x, the prediction is conducted by

$$C_{av}(x) = \text{AVN}(x_a, x_{v^{lip}}), \tag{5}$$

where x_a and $x_{v^{lip}}$ are the audio track and the lip-image sequence of x, respectively. As with the video-only and audio-only networks, in the inference phase of our ensemble model, either the single-vector representation or the predicted score corresponding to the "fake" of the audio-visual network is used by the DM module.

In addition to the above AV-HuBERT-based AVN setup, we also study another setup. The audio and video embeddings are extracted via AST (same as AN) and ViViT (same as VN). The corresponding cls tokens are concatenated as E_{av}. The prediction is also conducted by Eq. (5), but using (x_a, x_v) rather than (x_a, $x_{v^{lip}}$) as input. An overview of the AVTENet model based on this AVN setting is shown in Fig. 5. Compared with the AVTENet model in Fig. 3, the only difference is that the AV-HuBERT-based AVN is replaced by the AST&ViViT-based AVN.

D. Decision-Making Module (DM):

As shown in Fig. 3, the DM module of AVTENet is used to integrate the outputs of VN, AN, and AVN (trained separately as described previously), to produce final predictions that indicate potential forgery in the video. In this study, we investigate several fusion strategies.

1) Majority Voting: As shown in Fig. 6(a) and Algorithm 1, the DM module outputs "fake" if at least two of VN, AN, and AVN output "fake". The corresponding AVTENet model is termed AVTENet$_{mv}$. No additional models need to be trained for majority voting fusion.

2) Average Score Fusion: As shown in Fig. 6(b), the DM module outputs "fake" if the average of the output scores of VN, AN, and AVN exceeds a preset threshold. The

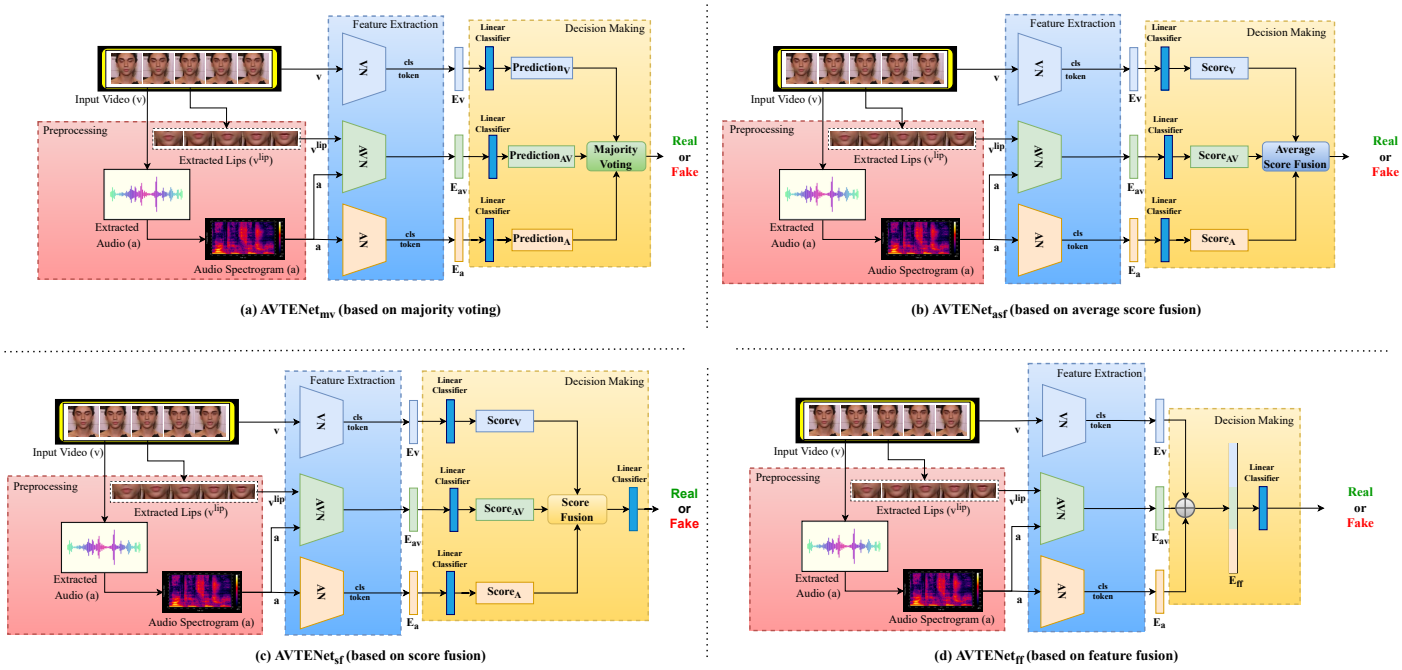


Fig. 6. Four variants of the proposed AVTENet model for video forgery detection based on different ensemble strategies: (a) $AVTENet_{mv}$ based on majority voting, (b) $AVTENet_{asf}$ based on average score fusion, (c) $AVTENet_{sf}$ based on score fusion, and (d) $AVTENet_{ff}$ based on feature fusion.

corresponding AVTENet model is termed AVTENet$_{asf}$. No additional models need to be trained for average score fusion.

3) Score Fusion: As shown in Fig. 6(c), in score fusion, the output scores of VN, AN, and AVN are concatenated as input to a linear layer trained on the same training data as the three component classifiers with the cross-entropy loss. When training the linear layer, the parameters of VN, AN, and AVN are fixed. The corresponding AVTENet model is termed AVTENet$_{sf}$.

4) Feature Fusion: As shown in Fig. 6(d), in feature fusion, the output representations of the penultimate layers of VN, AN, and AVN are concatenated as input to a linear layer trained on the same training data as the three component classifiers with the cross-entropy loss. When training the linear layer, the parameters of VN, AN, and AVN are fixed. The corresponding AVTENet model is termed AVTENet$_{ff}$.

IV. Experimental Results

This section presents the experiment setup, including the dataset, data analysis and preprocessing, evaluation metrics, training hyperparameters, and experimental results.

Algorithm 1 Majority Voting Algorithm for AVTENet$_{mv}$

Require: P_v, P_a, and P_{av} ($P_i \in 0$ or 1)
Ensure: AVTENet_Prediction P ($P \in 0$ or 1)
1: Initialize $vote_count$ variable, $vote_count = P_v + P_a + P_{av}$
2: **if** $vote_count \geq 2$ **then**
3: $P = 1$
4: **else**
5: $P = 0$
6: **end if**
7: **return** P

A. Dataset

Our experiments are conducted on the FakeAVCeleb dataset[1] [25]. As described in Sec. II-C, it contains 500 videos featuring 500 celebrities and 19,500 fake videos manipulated from these 500 videos. The videos corresponding to 70 celebrities are used for testing, and the videos corresponding to the remaining 430 celebrities are used for training. These videos are divided into four categories according to whether the audio and video modalities are manipulated, namely RealVideo-RealAudio (R_vR_a), RealVideo-FakeAudio (R_vF_a), FakeVideo-RealAudio (F_vR_a) and FakeVideo-FakeAudio (F_vF_a), where each video is carefully labeled to indicate its authenticity and facilitate training and testing forgery detection models. Fig. 7 shows a few samples of each category from the FakeAVCeleb dataset.

1) Testing Sets: To conduct a comprehensive evaluation of various unimodal (AN and VN) and multimodal (AVN and AVTENet) detection methods using the FakeAVCeleb dataset, we used eight different test sets, including Testset-I, Testset-II, faceswap, faceswap-wav2lip, fsgan, fsgan-wav2lip, RTVC, and wav2lip. Except for Testset-I and Testset-II, each test set contains the same set of 70 genuine videos of 70 subjects unseen in training and a different set of 70 fake videos generated using specific manipulation techniques. Testset-I and Testset-II are the main evaluation test sets, which also contain the above 70 genuine videos. However, for the 70 fake videos, Testset-I contains the same number of fake samples from each manipulation technique, while Testset-II contains an equal number of fake samples from the R_vF_a, F_vF_a and F_vR_a categories. Table I summarizes the manipulation modalities involved in each test set. Multiple test sets can provide an

[1]We do not use the DFDC dataset for two reasons. First, the subject's face in some videos may not be facing the camera, and our model needs to extract lip images. Second, its labeling lacks separate labels for acoustic and visual modalities in videos.

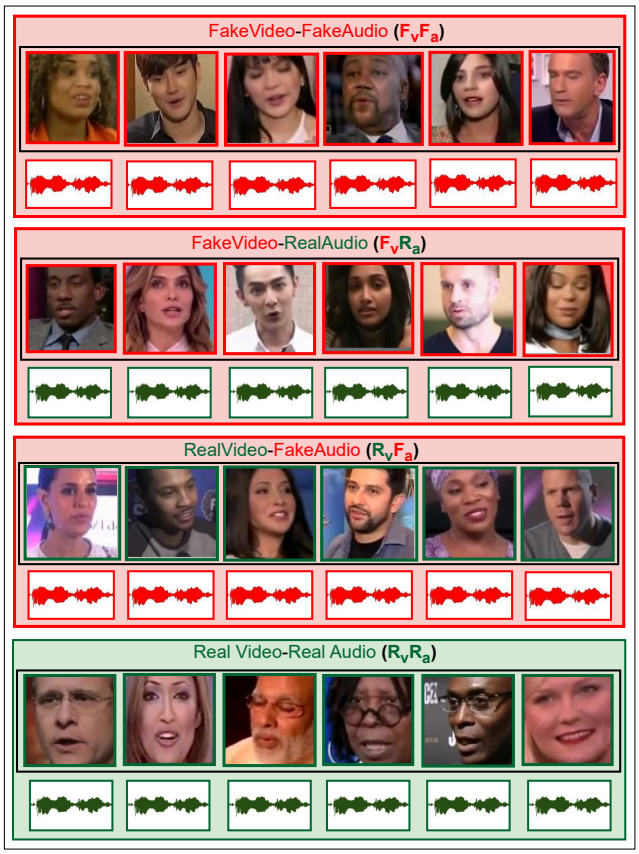


Fig. 7. Illustrations of the audio and visual streams of some samples selected from the FakeAVCeleb Dataset, where real streams are indicated in green, and fake ones are indicated in red.

exhaustive understanding of the strengths and limitations of AVTENet and individual classifiers on fake video samples generated using different deepfake techniques.

TABLE I
MANIPULATION MODALITIES IN EACH TEST SET. ✓ INDICATES MANIPULATION, WHILE X INDICATES NO MANIPULATION.

Testsets	Audio Manipulation	Video Manipulation
Testset-1	✓	✓
Testset-II	✓	✓
faceswap	X	✓
faceswap-wav2lip	✓	✓
fsgan	X	✓
fsgan-wav2lip	✓	✓
RTVC	✓	X
wav2lip	✓	✓

2) Training Sets for Different Models: In the experiments, the training data are video samples corresponding to 430 subjects in FakeAVCeleb. However, training unimodal and multimodal classifiers requires specific data settings, so we customize the dataset according to the requirements of each network. Table II shows the training data setting for each network according to the categorical labels of FakeAVCeleb. Table III shows the statistics of the real and fake samples used to train each network.

Video-only Network (VN): The VN classifier is trained to exploit manipulations in visual content. When constructing the training set for VN, as shown in Table II, the fake class includes training samples in the F_vF_a and F_vR_a categories because the visual content in these categories is manipulated.

TABLE II
TRAINING DATA SETTINGS FOR DIFFERENT NETWORKS.

Classifier	Class	Category
VN	Fake	F_vF_a, F_vR_a
	Real	R_vF_a, R_vR_a
AN	Fake	R_vF_a, F_vF_a
	Real	F_vR_a, R_vR_a
AVN	Fake	F_vF_a, F_vR_a, R_vF_a
	Real	R_vR_a

In contrast, the real class includes samples from the R_vF_a and R_vR_a categories, whose visual content is not manipulated. As shown in Table III, this setting results in the VN training set containing 17,809 video samples per class, of which the fake class contains 9,411 samples from F_vF_a and 8,398 samples from F_vR_a, while the real class contains 430 samples from R_vR_a, 430 samples from R_vF_a, and 16,949 samples from the external VoxCeleb1 dataset [68]. External data is used to balance the training data across classes.

TABLE III
NUMBER OF TRAINING SAMPLES FOR DIFFERENT NETWORKS.

Classifier	Class	Samples	Total Training Samples
VN	Fake	17,809	35,618
	Real	17,809	
AN	Fake	9,841	18,669
	Real	8,828	
AVN	Fake	18,239	36,411
	Real	18,172	

Audio-only Network (AN): AN learns to detect acoustic manipulations in videos. Therefore, the training set for the AN classifier is organized by treating categories with manipulated acoustic streams as fake and categories with unmanipulated acoustic streams as real. As shown in Table II, the fake class contains all samples of R_vF_a and F_vF_a, while the real class contains samples from F_vR_a and R_vR_a. As shown in Table III, the training set for AN contains 9,841 samples in the fake class and 8,828 samples in the real class. For the fake class, 430 samples belong to R_vF_a and 9,411 samples belong to F_vF_a, while for the real class, 8,398 samples belong to F_vR_a and 430 samples belong to R_vR_a.

Audio-Visual Network (AVN): As shown in Table II, for the AVN classifier, we specify F_vF_a, F_vR_a and R_vF_a categories, where acoustic and/or visual streams are manipulated, to the fake class. However, the real class only contains video samples from the R_vR_a category without acoustic and visual manipulations. As shown in Table III, the AVN training set contains 36,411 samples: 18,239 samples for the fake class and 18,172 for the real class. For the fake class, 9,411 samples belong to F_vF_a, 8,398 samples belong to F_vR_a, and 430 samples belong to R_vF_a. For the real class, in addition to the 430 samples from R_vR_a, we borrow 17,742 samples from the VoxCeleb1 dataset to balance the training data.

3) Data Analysis and Preprocessing: Each classifier involves different data processing steps to ensure effective training and evaluation. The video-only network takes as input a short clip of video; therefore, we split the video into multiple short clips and extract a sequence of consecutive frames from each clip to feed into the network. To prepare the data for the audio-only network, we extract the audio track with a sampling rate of 16kHz from each video. Furthermore, we

TABLE IV
DETECTION RESULTS OF THE VIDEO-ONLY NETWORK.

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.87	0.87	0.87	0.87
	Fake	0.87	0.87	0.87	
Testset-II	Real	0.76	0.87	0.81	0.80
	Fake	0.85	0.73	0.78	
faceswap	Real	1.00	0.87	0.93	0.94
	Fake	0.89	1.00	0.94	
faceswap-wav2lip	Real	1.00	0.87	0.93	0.94
	Fake	0.89	1.00	0.94	
fsgan	Real	0.95	0.87	0.91	0.91
	Fake	0.88	0.96	0.92	
fsgan-wav2lip	Real	1.00	0.87	0.93	0.94
	Fake	0.89	1.00	0.94	
RTVC	Real	0.50	0.87	0.64	0.50
	Fake	0.50	0.13	0.20	
wav2lip	Real	0.97	0.87	0.92	0.92
	Fake	0.88	0.97	0.93	

extract mel-spectrograms from these audios, which are then used as input to the audio-only network. We also extract log filterbank energy features that are used as one of the inputs to the audio-visual network. For the AV-HuBERT-based AVN, we extract the lip image frames from each video and pair them with the corresponding melspectral features as input to the network.

B. Hyperparameters

We use the Adam optimizer to train each classifier. AN (based on AST) is trained with a learning rate of 0.00001, VN (based on ViViT) is trained with a learning rate of 0.0001, and AVN (with AV-HuBERT) is trained with a learning rate of 0.002. Similarly, AVN (with AST&ViViT) is trained with a learning rate of 0.0001. In AVTENet, the linear layers for combining AST-based AN, ViViT-based VN, and AVN (both AVHuBERT-based and AST&ViViT-based) are trained with the learning rate of 0.002.

C. Evaluation Metrics

We use accuracy, precision, recall, and F1-score to evaluate the performance of individual unimodal/multimodal classifiers and our AVTENet classifier. They are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{9}$$

where TP, TN, FP, and FN denote True Positive (fake videos correctly detected as fake), True Negative (real videos correctly detected as real), False Positive (real videos incorrectly detected as fake), and False Negative (fake videos incorrectly detected as real), respectively. For all metrics, a higher value indicates better performance.

D. Experimental Results

This section provides detailed experimental results and analysis. Testset-I and Test-II are the main evaluation test sets of the FakeAVCeleb dataset.

1) Results of the Video-only Network: The results of the VN classifier on eight test sets are shown in Table IV. We observe that the VN classifier performs better on the test sets with visual manipulations because it utilizes visual features to detect forgeries in videos. The accuracy on the faceswap, faceswap-wav2lip, fsgan, fsgan-wav2lip, and wav2lip test sets are all above 0.90, except for RTVC, which achieves an accuracy of 0.50. Table I shows that, except for RTVC, all these test sets contain visual manipulations, making VN perform better on these test sets. The poor performance on RTVC is due to the lack of visual manipulations. Notably, both faceswap-wav2lip and fsgan-wav2lip include acoustic manipulations as well, but the VN classifier only detects visual manipulations in these cases. Furthermore, the accuracy of the two main test sets (0.87 for Test set-I and 0.80 for Testset-II) is relatively low compared to other test sets. This is because, for both test sets, the fake class contains videos with or without visual manipulation. In conclusion, the VN classifier performs better when the video is visually manipulated, but fails when the video only contains acoustic manipulation.

2) Results of the Audio-only Network: The AN classifier is a unimodal method that uses acoustic features to detect whether a video is real or fake. The results of AN on eight test sets are shown in Table V. We observe a similar pattern to the results of the VN classifier in Table IV. The detection accuracy is almost perfect on the faceswap-wav2lip, fsgan-wav2lip, and RTVC test sets (all containing acoustic manipulation in the fake class). However, AN fails on the faceswap and fsgan test sets, where fake instances are only visually manipulated while the audio tracks are all real. Furthermore, the detection accuracies of Testset-I and Testset-II (0.80 and 0.84, respectively) are lower than the strong performance of the other test sets. This is because, for both test sets, the fake class contains videos with or without acoustic manipulation. The results how that the AN classifier can effectively detect acoustically manipulated videos, but cannot detect fake videos containing only visual manipulations.

3) Results of the Audio-Visual Network: Compared to the VN and AN classifiers, which can only detect visual and acoustic manipulations in fake videos, respectively, the AVN classifier utilizes both acoustic and visual information to detect forgeries in videos. Table VI shows the results of the AV-HuBERT-based AVN classifier on eight test sets. Several observations can be drawn from the table. First, the AVN

TABLE V
DETECTION RESULTS OF THE AUDIO-ONLY NETWORK.

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.71	1.00	0.83	0.80
	Fake	1.00	0.60	0.75	
Testset-II	Real	0.76	1.00	0.86	0.84
	Fake	1.00	0.69	0.81	
faceswap	Real	0.50	1.00	0.67	0.50
	Fake	0.00	0.00	0.00	
faceswap-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
fsgan	Real	0.50	1.00	0.67	0.50
	Fake	0.00	0.00	0.00	
fsgan-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
RTVC	Real	0.97	1.00	0.99	0.99
	Fake	1.00	0.97	0.99	
wav2lip	Real	0.69	1.00	0.82	0.78
	Fake	1.00	0.56	0.72	

TABLE VI
DETECTION RESULTS OF THE AV-HUBERT-BASED AUDIO-VISUAL NETWORK.

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.92	0.96	0.94	0.94
	Fake	0.96	0.91	0.93	
Testset-II	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	
faceswap	Real	0.79	0.96	0.86	0.85
	Fake	0.95	0.74	0.83	
faceswap-wav2lip	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	
fsgan	Real	0.91	0.96	0.93	0.93
	Fake	0.95	0.90	0.93	
fsgan-wav2lip	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	
RTVC	Real	0.96	0.96	0.96	0.96
	Fake	0.96	0.96	0.96	
wav2lip	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	

classifier achieves good detection performance on all test sets except the faceswap test set. One possible reason is that the faceswap test set only involves visual manipulation, and the AV-HuBERT feature extractor only extracts visual features from lip images, thereby losing information outside the lip region. Second, although the AV-HuBERT-based AVN classifier performs well on the test sets containing manipulation in a single modality . However, its performance may be slightly lower than audio-only and video-only classifiers focusing on one modality. For instance, the AVN classifier achieves an accuracy of 0.96 on RTVC (fake samples containing only acoustic manipulation), but the AN classifier achieves an accuracy of 0.99 on the same test set (see Table V). This is because the AN classifier only focuses on the acoustic modality, while the AVN classifier will inevitably be interfered with by the visual feature extracted from the real visual modality when detecting acoustic manipulation. Third, the AV-HuBERT-based AVN classifier achieves accuracy of 0.94 and 0.98 on Testset-I and Testset-II, whose fake instances contain various manipulations in acoustic and visual streams, outperforming the VN classifier (cf. 0.87 and 0.80 in Table IV) and the AN classifier (cf. 0.80 and 0.84 in Table V). The results reveal that the AVN classifier successfully detects video forgeries by considering both acoustic and visual information. It generally outperforms VN and AN classifiers, which focus on only visual or acoustic manipulation.

4) Results of AVTENet: AVTENet is an ensemble of the above VN, AN, and AVN networks. We compare four fusion strategies, namely majority voting (mv), average score fusion (asf), score fusion (sf), and feature fusion (ff). The resulting models are denoted as $AVTENet_{mv}$, $AVTENet_{asf}$, $AVTENet_{sf}$, and $AVTENet_{ff}$, respectively. AVN is based on AV-HuBERT.

AVTENet$_{\mathbf{mv}}$: Table VII shows the results of $AVTENet_{mv}$ on eight different test sets. $AVTENet_{mv}$ achieves perfect accuracy of 1.00 on faceswap-wav2lip, fsgan-wav2lip, and Testset-II and high accuracy of 0.99, 0.98, and 0.96 on wav2lip, RTVC, and Testset-I. This result confirms the advantages of integrating VN, AN, and AVN networks in fully considering visual and acoustic information in fake video detection. However, $AVTENet_{mv}$ achieves relatively low performance on faceswap and fsgan. As explained before, this may be because the faceswap and fsgan test sets only involve visual manipulations, while the AV-HuBERT feature extractor ignores visual information outside the lip region. Comparing Table VII with Tables IV, V, and VI, $AVTENet_{mv}$ outperforms all its component networks in all test conditions except for VN on faceswap and AN on RTVC. This result shows that the ensemble $AVTENet_{mv}$ comes with some tradeoffs in fake detection compared to modality-specific classifiers operating on the corresponding unimodal manipulation.

AVTENet$_{\mathbf{asf}}$: Table VIII shows the results of $AVTENet_{asf}$ on eight different test sets. Comparing Table VIII and Table VII, we can see that $AVTENet_{asf}$ performs almost the same as $AVTENet_{mv}$, except that the former achieves a slightly lower accuracy than the latter on the faceswap test set (0.86 vs 0.87).

AVTENet$_{\mathbf{sf}}$: Table IX shows the results of $AVTENet_{sf}$ on eight different test sets. Comparing Table IX with Tables VII and VIII, we can see that $AVTENet_{sf}$ performs more stably

TABLE VII
DETECTION RESULTS OF $AVTENet_{mv}$ (MAJORITY VOTING).

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.92	1.00	0.96	0.96
	Fake	1.00	0.91	0.96	
Testset-II	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
faceswap	Real	0.80	1.00	0.89	0.87
	Fake	1.00	0.74	0.85	
faceswap-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
fsgan	Real	0.88	1.00	0.93	0.93
	Fake	1.00	0.86	0.92	
fsgan-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
RTVC	Real	0.96	1.00	0.98	0.98
	Fake	1.00	0.96	0.98	
wav2lip	Real	0.99	1.00	0.99	0.99
	Fake	1.00	0.99	0.99	

TABLE VIII
DETECTION RESULTS OF $AVTENet_{asf}$ (AVERAGE SCORE FUSION).

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.92	1.00	0.96	0.96
	Fake	1.00	0.91	0.96	
Testset-II	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
faceswap	Real	0.79	1.00	0.88	0.86
	Fake	1.00	0.73	0.84	
faceswap-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
fsgan	Real	0.88	1.00	0.93	0.93
	Fake	1.00	0.86	0.92	
fsgan-wav2lip	Real	1.00	1.00	1.00	1.00
	Fake	1.00	1.00	1.00	
RTVC	Real	0.96	1.00	0.98	0.98
	Fake	1.00	0.96	0.98	
wav2lip	Real	1.00	0.99	0.99	0.99
	Fake	0.99	1.00	0.99	

TABLE IX
DETECTION RESULTS OF $AVTENet_{sf}$ (SCORE FUSION).

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.96	0.91	0.93	0.94
	Fake	0.92	0.96	0.94	
Testset-II	Real	1.00	0.91	0.96	0.96
	Fake	0.92	1.00	0.96	
faceswap	Real	0.93	0.91	0.92	0.92
	Fake	0.92	0.93	0.92	
faceswap-wav2lip	Real	1.00	0.91	0.96	0.96
	Fake	0.92	1.00	0.96	
fsgan	Real	0.93	0.91	0.92	0.92
	Fake	0.92	0.93	0.92	
fsgan-wav2lip	Real	1.00	0.91	0.96	0.96
	Fake	0.92	1.00	0.96	
RTVC	Real	0.96	0.91	0.93	0.94
	Fake	0.92	0.96	0.94	
wav2lip	Real	1.00	0.91	0.96	0.96
	Fake	0.92	1.00	0.96	

TABLE X
DETECTION RESULTS OF $AVTENet_{ff}$ (FEATURE FUSION).

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset-I	Real	0.96	0.97	0.96	0.96
	Fake	0.97	0.96	0.96	
Testset-II	Real	1.00	0.97	0.99	0.99
	Fake	0.97	1.00	0.99	
faceswap	Real	0.93	0.97	0.95	0.95
	Fake	0.97	0.93	0.95	
faceswap-wav2lip	Real	1.00	0.97	0.99	0.99
	Fake	0.97	1.00	0.99	
fsgan	Real	0.93	0.97	0.95	0.95
	Fake	0.97	0.93	0.95	
fsgan-wav2lip	Real	1.00	0.97	0.99	0.99
	Fake	0.97	1.00	0.99	
RTVC	Real	0.96	0.97	0.96	0.96
	Fake	0.97	0.96	0.96	
wav2lip	Real	1.00	0.97	0.99	0.99
	Fake	0.97	1.00	0.99	

than $AVTENet_{mv}$ and $AVTENet_{asf}$ across test sets. The stability may be attributed to the additional linear layer taking as input the output scores of the VN, AN, and AVN networks, as it learns to balance the contributions of these three component networks. However, although $AVTENet_{sf}$ outperforms $AVTENet_{mv}$ and $AVTENet_{asf}$ on the faceswap test set (0.92 vs 0.87 and 0.86), it performs worse than $AVTENet_{mv}$ and $AVTENet_{asf}$ on other test sets.

$\text{AVTENet}_{\mathbf{ff}}$: Table X shows the results of $AVTENet_{ff}$ on eight different test sets. We can see that $AVTENet_{ff}$ performs best among four ensemble models on faceswap, fsgan, and Testset-II, and achieves near-perfect performance on faceswap-wav2lip, fsgan-wav2lip, wav2lip, and Testset-II. While other ensemble models may perform well on specific test sets, $AVTENet_{ff}$ maintains stable and reliable performance on all eight test sets. The good performance may be attributed to the additional linear layers that takes as input the embeddings of the penultimate layers of the VN, AN, and AVN networks, as it learns to balance the contributions of these three component networks. As shown in Figs. 8 and 9, the discriminative ability of the fused feature $\mathbf{E_{ff}}$ is further improved in the additional linear classification layer of $AVTENet_{ff}$. In summary, feature fusion is the best fusion strategy in this study.

5) Comparison of The Above Models: Fig. 10 depicts the accuracy of the above seven models on the two main test sets, Testset-I and Testset-II. The video-only model (VN) is good at identifying visual manipulations in videos. Likewise, the audio-only model (AN) is good at identifying manipulated

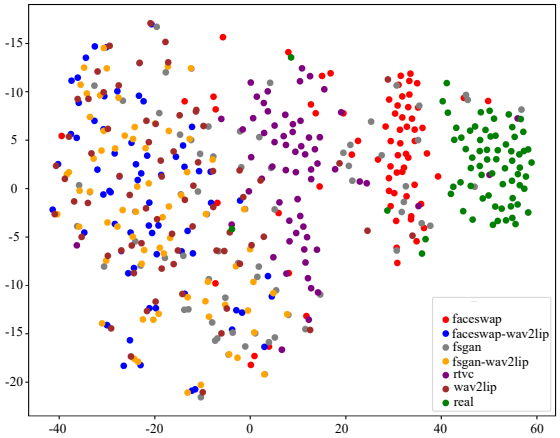


Fig. 8. t-SNE visualization of the fused feature $\mathbf{E_{ff}}$ in $AVTENet_{ff}$.

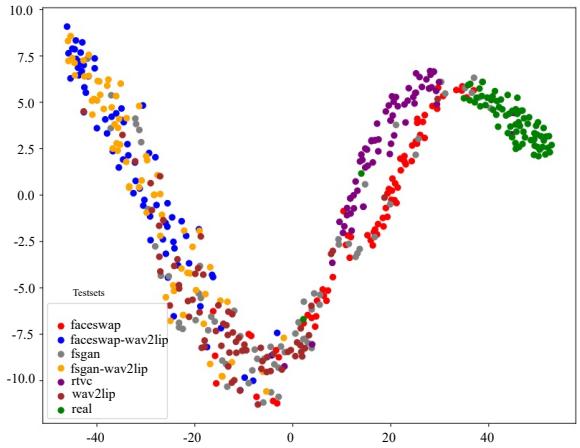


Fig. 9. t-SNE visualization of features in the linear classifier of $AVTENet_{ff}$.

acoustic content in videos. But they only focus on a single modality. In contrast to VN and AN, the AVN and all variants of AVTENet can identify acoustic and visual manipulations in videos. Although $AVTENet_{mv}$ and $AVTENet_{asf}$ have higher accuracy than $AVTENet_{ff}$ on Testset-II, the latter shows higher stability in detecting various manipulation types in videos.

Comparing Tables VII-IX with Table IV, we can see that among the variants of AVTENet, only $AVTENet_{ff}$ can outperform the VN model on all test sets, while the remaining AVTENet models perform worse than the VN model on the faceswap test set (fake samples containing only visual manipulation), despite outperforming the VN model on the other test sets. This is because the VN model only focuses on the visual modality, while the AVTENet models may be interfered with by the acoustic feature extracted from the real auditory modality when detecting visual manipulation. Furthermore, comparing Tables VII-IX with Table V, all AVTENet models outperform the AN model on all test sets except RTVC (fake samples containing only acoustic manipulation). This happens because the AN model only focuses on the acoustic modality, while the AVTENet models may be interfered with by the visual feature extracted from the real visual modality when detecting acoustic manipulation.

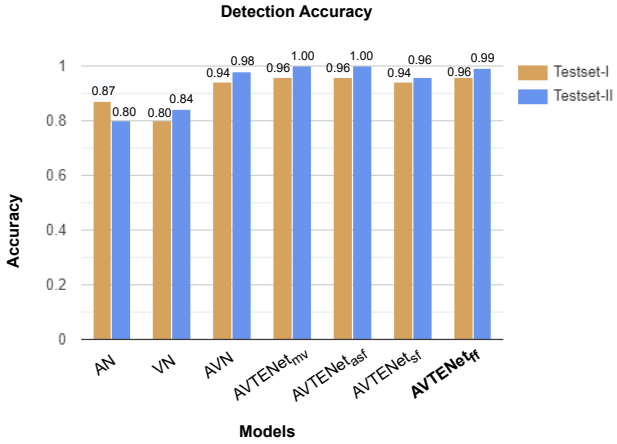


Fig. 10. Performance comparison of our models on two main test sets, Testset-I and Testset-II.

TABLE XI
PERFORMANCE COMPARISON OF AVTENET (AST_VIVIT_AST&VIVIT) AND AVTENET (AST_VIVIT_AV-HUBERT).

Models	Ensemble Strategies	faceswap	faceswap-wav2lip	fsgan	fsgan-wav2lip	RTVC	wav2lip	Testset-I	Testset-II
AVTENet (AST_ViViT_AST&ViViT)	$AVTENet_{mv}$	**0.96**	0.99	0.94	0.99	0.61	0.98	0.87	0.94
	$AVTENet_{asf}$	**0.96**	0.99	0.93	0.99	0.63	0.98	0.88	0.94
	$AVTENet_{sf}$	0.91	0.93	0.88	0.93	0.52	0.91	0.80	0.87
	$AVTENet_{ff}$	0.92	0.97	0.88	0.96	0.52	0.90	0.80	0.87
AVTENet (AST_ViViT_AV-HuBERT)	$AVTENet_{mv}$	0.87	**1.00**	0.93	**1.00**	**0.98**	**0.99**	**0.96**	**1.00**
	$AVTENet_{asf}$	0.86	**1.00**	0.93	**1.00**	**0.98**	**0.99**	**0.96**	**1.00**
	$AVTENet_{sf}$	0.92	0.96	0.92	0.96	0.94	0.96	0.94	0.96
	$AVTENet_{ff}$	0.95	0.99	**0.95**	0.99	0.96	**0.99**	**0.96**	0.99

6) Comparison of AVTENet with Different AVNs: Our proposed AVTENet model is an ensemble of an AST-based AN model, a ViViT-based VN model, and an AV-HuBERT-based AVN model. Although $AVTENet_{ff}$ performs well on all test sets, its performances on the faceswap and fsgan test sets are relatively low, where the fake instances are only visually manipulated while audio tracks are real. We speculate that the reason may be that the AV-HuBERT feature extractor in AVN only extracts visual information from lip images and ignores information outside the lip region. Therefore, we also implement another AVN based on AST and ViViT. Here, we compare AVTENet (AST_ViViT_AV-HuBERT) and AVTENet (AST_ViViT_AST&ViViT). From Table XI, although the AST_ViViT_AST&ViViT-based $AVTENet_{mv}$ and $AVTENet_{asf}$ models do slightly improve the performance on the faceswap test set, the AST_ViViT_AST&ViViT-based AVTENet models perform worse than the AST_ViViT_AV-HuBERT-based AVTENet models. The AST and ViViT models are already used in the AN and VN models. Therefore, they cannot provide additional information on the AVN model. Surprisingly, the AST_ViViT_AST&ViViT-based AVTENet models almost fail on the RTVC test set. To fully utilize face images in the AVN model, we may need to apply new pre-trained visual models.

7) Comparison of AVTENet and Other Models: Finally, as shown in Fig. 11 and Table XII, we conduct a thorough comparison of our proposed $AVTENet_{ff}$ model with various existing unimodal, multimodal, fusion, and ensemble methods. Several unimodal, multimodal and ensemble methods were presented in [69], but most of them failed to detect acoustic and visual manipulations in videos well. VGG16, a unimodal approach exclusively trained on visual data, outperforms all other unimodal approaches reported in [69] for video forgery detection with an accuracy of 0.81. Xception, another unimodal method reported in [69], is trained exclusively on the audio modality and achieves an accuracy of only 0.76.

TABLE XII
PERFORMANCE COMPARISON OF OUR ENSEMBLE MODEL AND SEVERAL EXISTING UNIMODAL, MULTIMODAL, FUSION, AND ENSEMBLE MODELS. DFD IN THE FIRST COLUMN REFERS TO THE DEEPFAKE DETECTION METHOD. "V", "A" AND "AV" STAND FOR VISUAL, AUDIO AND AUDIO-VISUAL MODALITIES, RESPECTIVELY.

DFD Method	Model	Modality	Class	Precision	Recall	F1-score	Accuracy
Unimodal [69]	VGG16	V	Real	0.69	0.90	0.78	0.81
			Fake	0.87	0.78	0.82	
Unimodal [69]	Xception	A	Real	0.88	0.61	0.72	0.76
			Fake	0.70	0.91	0.80	
Unimodal [65]	LipForensics	V	Real	0.70	0.91	0.80	0.76
			Fake	0.88	0.61	0.72	
Ensemble (Soft-Voting) [69]	VGG16	AV	Real	0.69	0.90	0.78	0.78
			Fake	0.90	0.69	0.78	
Ensemble (Hard-Voting) [69]	VGG16	AV	Real	0.69	0.90	0.78	0.78
			Fake	0.90	0.69	0.78	
Multimodal-1 [69]	Multimodal-1	AV	Real	0.00	0.00	0.00	0.50
			Fake	0.50	1.00	0.66	
Multimodal-2 [69]	Multimodal-2	AV	Real	0.71	0.59	0.64	0.67
			Fake	0.65	0.76	0.70	
Multimodal-3 [69]	CDCN	AV	Real	0.50	0.07	0.12	0.52
			Fake	0.50	0.94	0.65	
Multimodal-4 [15]	Not-made-for-each-other	AV	Real	0.62	0.99	0.76	0.69
			Fake	0.94	0.40	0.57	
Multimodal [9]	Ensemble	AV	Real	0.83	0.99	0.90	0.89
			Fake	0.98	0.80	0.88	
Multimodal [10]	AV-Lip-Sync	AV	Real	0.93	0.96	0.94	0.94
			Fake	0.96	0.93	0.94	
Modality Mixing [70]	Multimodaltrace	AV	Real	-	-	-	0.93
			Fake	-	-	-	
Ensemble [71]	AVFakeNet	AV	Real	-	-	-	0.93
			Fake	-	-	-	
Fusion [72]	AVoiD-DF	AV	Real	-	-	-	0.84
			Fake	-	-	-	
Fusion [73]	MIS-AVioDD	AV	Real	-	-	-	0.96
			Fake	-	-	-	
Ensemble (ours)	$\mathbf{AVTENet_{ff}}$	AV	Real	**1.00**	**0.97**	**0.99**	**0.99**
			Fake	**0.97**	**1.00**	**0.99**	

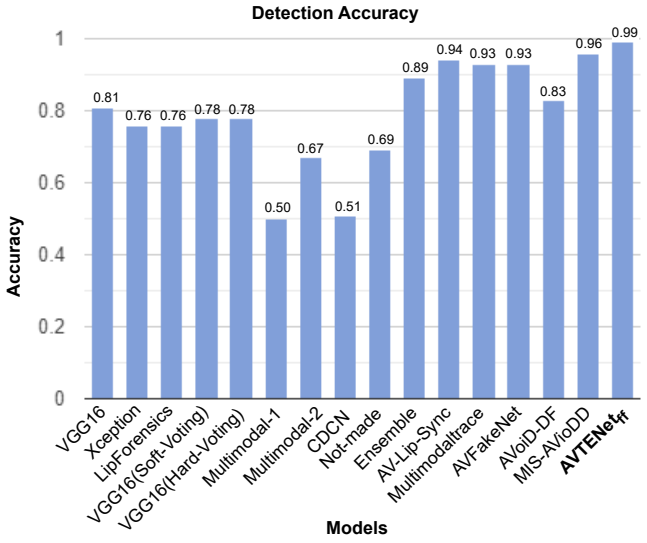


Fig. 11. The comparison of our proposed approach ($AVTENet_{ff}$) with other existing unimodal, multimodal, fusion, and ensemble benchmarks.

Similarly, Lip Forensics [42] exploits visual lip movements to detect forgeries in videos with an accuracy of 0.76. Although Ensemble (soft-voting and hard-voting), Multimodal-1, Multimodal-2, and Multimodal-3 utilize both audio and visual modalities for deepfake detection, their performance is still unsatisfactory. Not-made-for-each other [15] is an audio-visual dissonance-based approach that detects video forgeries with an accuracy of 0.69. The multimodal ensemble approach proposed in [9] utilizes three CNN-based classifiers to identify forgeries and achieves an accuracy of 0.89. Another audio-visual approach proposed in [10] successfully identifies fake videos using lip synchronization, and the model achieves an accuracy of 0.94. Moreover, the ensemble approach Multimodaltrace [70] achieves an accuracy of 0.93. AVFakeNet [71] is also an ensemble approach using audio and video transformer encoders, which achieves an accuracy of 0.93. AVoiD-DF [72] is another audio-visual fusion approach, but it only achieves an accuracy of 0.84. Furthermore, MIS-AVioDD [73] is also a fusion approach that jointly uses the modality invariant and modality-specific representations to detect audio-visual forgery and achieves an accuracy of 0.96. Table XII shows that our ensemble approach outperforms other existing models, with an accuracy of 0.99 on Testset-II of the FakeAVCeleb dataset, which is a new SOTA performance.

8) Comparison of AVTENet and Humans: We conducted a preliminary comparison of AVTENet's detection capabilities with human cognitive proficiency in detecting audio-visual deepfakes using a manually curated subset of 30 videos (15 real videos and 15 fake videos) from the FakeAVCeleb test set to ensure a diverse and challenging evaluation set, with all fake videos featuring audio-visual manipulation.

The videos were presented to 24 subjects in three phases to classify each video as real or fake. In the first phase (audio-only), participants judged whether the video was authentic or not by listening to the audio track alone. In the second phase (video-only), participants judged the authenticity of the video using only visual input. Finally, in the third phase (audio-visual), participants integrated the audio and visual streams to classify the video as real or fake. As shown in Table XIV, humans rely on their own perception and reasoning capabilities to achieve an accuracy of 0.58 under the audio-only condition, 0.63 under the video-only condition, and 0.76 under the audio-visual condition. From Table XIV, it is clear that, under three conditions, AST (audio-only network), ViViT (video-only network), and AVTENet (audio-visual network) outperform humans (0.77 vs 0.58, 0.83 vs 0.63, and 0.93 vs 0.76), respectively. The results highlight the importance of multimodal sensory integration for enhancing human and model performance in video forgery detection; especially for detection models, combining audio and visual cues can lead to significant performance improvements.

TABLE XIII
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AND OTHER MODELS TESTED ON THE DFDC TEST SET.

Model	AUC Score (%)
LipForensics [65]	53.10
CViT [74]	56.70
MesoNet [11]	41.20
MDS [15]	73.80
AVoiD-DF [72]	80.60
Proposed model ($AVTENet_{ff}$)	64.10
Proposed model ($AVTENet_{ff}$) (on the frontal face subset)	83.84

TABLE XIV
COMPARISON OF HUMANS AND MODELS IN VIDEO FORGERY DETECTION.

Humans		Models	
Modalities	**Accuracy**	**Modalities**	**Accuracy**
Audio-Only	0.58	AST (Audio-Only)	0.77
Video-Only	0.63	ViViT (Video-Only)	0.83
Audio-visual	0.76	AVTENet (Audio-visual)	0.93

9) Generalization Test: Generalization remains a significant challenge for video forgery detection tasks, including multimodal deepfake detection. To evaluate the generalization ability of our proposed method, we directly test the proposed model trained on the FakeAVCeleb dataset on the DFDC test set, and compare it with other models under the same setting. We extract the test videos and preprocess them following [15] and [75]. As shown in Table XIII, our model achieves an AUC (area under the receiver operating characteristic curve) score of 64.10%, which is better than the performance of LipForensics [65], CViT [74], and MesoNet [11], but worse than the performance of MDS [15] and AVoiD-DF [72]. The superior performance of these methods may be due to the fact that they focus on the entire image frame rather than just the frontal face or lip regions.

Our study focuses on developing a multimodal deepfake detection method for frontal face videos. However, the DFDC test set contains various types of deepfake videos, including distant faces, side faces, multiple speakers, illumination variations, and low-light conditions. This variation poses challenges to the accurate extraction of faces and lips. Therefore, we took the frontal face videos from the DFDC test set and created a frontal face video test subset. As shown in Table XIII, our model achieves an AUC score of 83.84% on the frontal face video test subset. The results show that our proposed model remains highly effective for frontal face videos where facial identity or expression as well as lip movements may be manipulated, and shows reasonable performance on the entire

test set. The results also show that our model has quite good generalization ability.

V. Limitations

This study focuses on the authenticity detection of frontal face videos. Lack of frontal face, extreme environmental conditions, illumination variation, and multiple speakers in videos pose challenges to the face detection and lip capture modules in our method. For such types of videos, complete frame-based image analysis is required. However, in the proposed ensemble approach, the video-only model analyzes visual inconsistencies in facial movements, lighting, and other video content, independent of lip image data. While utilizing lip image data, the audio-visual model supplements the other two models by cross-referencing audio-visual synchronization to improve detection accuracy. Furthermore, our method is bimodal, requiring the input to contain both visual and acoustic modalities. To train our model, bimodal manipulations are required to train independent modality-specific networks and integrate their features into the decision-making module.

Due to the lack of suitable audio-visual datasets, it is currently not possible to test the generalization ability of the proposed model on more datasets. In addition to awaiting new publicly available audio-visual deepfake datasets, we are also preparing to expand the DeepfakeTIMIT dataset [76], where the audio in both real and fake videos is consistently authentic. We aim to apply various cutting-edge voice conversion technologies for acoustic manipulation.

As our ensemble approach consists of multiple transformers, the overall computational cost is higher compared to CNN-based or single transformer-based models in the literature. The number of trainable parameters in our video-only, audio-visual, and audio-only networks are 114.75M, 129.31M, and 87.73M respectively. The number of parameters in the linear layers of the decision-making module is 2.89M. In total, our proposed AVTENET model contains 334.68M trainable parameters. For inference time, we evaluate our proposed model on an NVIDIA GeForce RTX 2080 Ti GPU. To ensure accuracy, we executed the proposed model in the evaluation mode and performed multiple warm-up iterations before measurement to eliminate cold start overhead. The average inference time per video sample is 0.55 seconds. The recorded inference time only includes the forward pass of the proposed model and does not include any pre- or post-processing time.

VI. Conclusions and Future Work

In this study, we demonstrated the effectiveness of ensemble learning and studied deepfake video detection from an audio-visual perspective. We proposed a novel audio-visual transformer-based ensemble learning solution for effective and scalable video forgery detection. This approach is particularly effective for videos containing acoustic and visual manipulations, which cannot rely on existing methods because they can only detect unimodal manipulations or lack training diversity. Specifically, we devised a transformer-based model with a powerful detection ability due to its sequence modeling, parallelization, and attention mechanisms. We illustrated the performances of our multimodal transformer-based ensemble model on several manipulation techniques and compared them with various existing unimodal, multimodal, fusion, and ensemble models. The proposed model achieved state-of-the-art performance on the FakeAVCeleb dataset. To enhance trust and security in digital media environments, our model can be used to instantly detect deepfakes on social media platforms, which aids in content moderation and forensic analysis. Furthermore, the proposed solution can be extended through various foundation models trained using large amounts of data.

Acknowledgments

This work was supported in part by Academia Sinica under Grant AS-GC-111-M01, and by the National Council of Science and Technology, Taiwan, under Grants NSTC 113-2634-F-002-003 and NSTC 112-2221-E-001-009-MY3.

References

[1] A. Ray, "Disinformation, deepfakes and democracies: The need for legislative reform," *The University of New South Wales Law Journal*, vol. 44, no. 3, pp. 983–1013, 2021.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations*, 2014.

[4] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.

[5] S. Scheliga, T. Kellermann, A. Lampert, R. Rolke, M. Spehr, and U. Habel, "Neural correlates of multisensory integration in the human brain: an ale meta-analysis," *Reviews in the Neurosciences*, vol. 34, no. 2, pp. 223–245, 2023.

[6] J. Lin, "Sensory inputs guiding cognitive behaviors and decision making," *Highlights in Science, Engineering and Technology*, vol. 74, pp. 1399–1404, 2023.

[7] S.-F. Zhang, J.-H. Zhai, B.-J. Xie, Y. Zhan, and X. Wang, "Multimodal representation learning: Advances, trends and challenges," in *Proceedings of the International Conference on Machine Learning and Cybernetics*, 2019, pp. 1–6.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

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

[10] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "Lip Sync Matters: A novel multimodal forgery detector," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 1885–1892.

[11] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a compact facial video forgery detection network," in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 6105–6114.

[14] X. Han, V. Morariu, P. I. Larry Davis *et al.*, "Two-stream neural networks for tampered face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 19–27.

[15] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 439–447.

[16] W. Ahmad, I. Ali, A. Shahzad, A. Hashmi, and F. Ghaffar, "ResViT: A framework for deepfake videos detection," *International Journal of Electrical and Computer Engineering Systems*, vol. 13, no. 9, pp. 807–813, 2022.

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[18] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proceedings of the Interspeech Conference*, 2021.

[19] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proceedings of the International Conference on Learning Representations*, 2021.

[20] L. Guo, Z. Gao, J. Qu, S. Zheng, R. Jiang, Y. Lu, and H. Qiao, "Transformer-based spiking neural networks for multimodal audio-visual classification," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 3, pp. 1077–1086, 2023.

[21] S. Kumar, A. Sur, and R. D. Baruah, "DatUS: Data-driven unsupervised semantic segmentation with pre-trained self-supervised vision transformer," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 5, pp. 1775–1788, 2024.

[22] R. Wang, W. Jo, D. Zhao, W. Wang, A. Gupte, B. Yang, G. Chen, and B.-C. Min, "Husformer: A multi-modal transformer for multi-modal human state recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 4, pp. 1374–1390, 2024.

[23] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.

[24] D.-K. Jeong, H.-G. Kim, and J. Y. Kim, "Automated video classification system driven by characteristics of emotional human brainwaves caused by audiovisual stimuli," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 651–661, 2022.

[25] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[26] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[27] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9459–9468.

[28] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2019.

[29] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–14, 2018.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[31] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5933–5942.

[32] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.

[33] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm, "Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1022.

[34] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[35] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8261–8265.

[36] W. Wu, W. Zhou, W. Zhang, H. Fang, and N. Yu, "Capturing the lighting inconsistency for deepfake detection," in *Proceedings of the International Conference on Artificial Intelligence and Security*, 2022, pp. 637–647.

[37] D.-K. Kim and K.-S. Kim, "Generalized facial manipulation detection with edge region feature extraction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2828–2838.

[38] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[39] A. Borji, "Qualitative failures of image generation models and their application in detecting deepfakes," *Image and Vision Computing*, vol. 137, p. 104771, 2023.

[40] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings of the IEEE Winter Applications of Computer Vision Workshops*, 2019, pp. 83–92.

[41] E. Prashnani, M. Goebel, and B. Manjunath, "Generalizable deepfake detection with phase-based motion analysis," *arXiv preprint arXiv:2211.09363*, 2022.

[42] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2020, pp. 1–6.

[43] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 1821–1828.

[44] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid transformer network for deepfake detection," in *Proceedings of the International Conference on Content-based Multimedia Indexing*, 2022, pp. 8–14.

[45] N. Gautam and D. K. Vishwakarma, "Obscenity detection in videos through a sequential convnet pipeline classifier," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 310–318, 2022.

[46] J. R. Williams, Y. A. Markov, N. A. Tiurina, and V. S. Störmer, "What You See Is What You Hear: Sounds alter the contents of visual perception," *Psychological Science*, vol. 33, no. 12, pp. 2109–2122, 2022.

[47] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2823–2832.

[48] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 800–14 809.

[49] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 943–952.

[50] S. Muppalla, S. Jia, and S. Lyu, "Integrating audio-visual features for multimodal deepfake detection," *arXiv preprint arXiv:2310.03827*, 2023.

[51] S. Asha, P. Vinod, I. Amerini, and V. G. Menon, "A novel deepfake detection framework using audio-video-textual features," *Research Square*, 2022.

[52] H. Zhao, W. Zhou, D. Chen, W. Zhang, and N. Yu, "Self-supervised transformer for deepfake detection," *arXiv preprint arXiv:2203.01265*, 2022.

[53] C.-S. Sung, J.-C. Chen, and C.-S. Chen, "Hearing and Seeing Abnormality: Self-supervised audio-visual mutual learning for deepfake detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[54] W. Liu, X. Wei, T. Lei, X. Wang, H. Meng, and A. K. Nandi, "Data-fusion-based two-stage cascade framework for multimodality face anti-spoofing," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 672–683, 2021.

[55] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "LieNet: A deep convolution neural network framework for detecting deception," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 971–984, 2021.

[56] C. Yuan, S. Jiao, X. Sun, and Q. J. Wu, "MFFFLD: A multimodal-feature-fusion-based fingerprint liveness detection," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 648–661, 2021.

[57] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.

[58] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[59] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of the Interspeech*, 2018.

[60] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 3677–3685.

[61] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7184–7193.

[62] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 484–492.

[63] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of the Advances in Neural Information Processing Systems*, 2018.

[64] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[65] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

[66] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 6319–6323.

[67] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[68] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proceedings of the Interspeech Conference*, 2017.

[69] 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*, 2021, pp. 7–15.

[70] 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*, 2023, pp. 993–1000.

[71] 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.

[72] 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.

[73] V. S. Katamneni and A. Rattani, "MIS-AVioDD: Modality invariant and specific representation for audio-visual deepfake detection," *arXiv preprint arXiv:2310.02234*, 2023.

[74] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[75] M. Astrid, E. Ghorbel, and D. Aouada, "Statistics-aware audio-visual deepfake detector," in *Proceedings of the IEEE International Conference on Image Processing*, 2024.

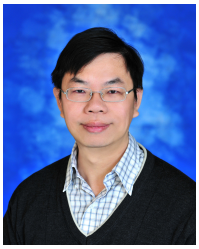
[76] P. Korshunov and S. Marcel, "DeepFakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.



Ammarah Hashmi received her B.S. in Software Engineering from COMSATS University, Pakistan in 2017, followed by an M.S. in Software Engineering from Beijing Institute of Technology, China in 2019, where she specialized in gamification and its applications in user engagement and digital experiences. She is currently pursuing her Ph.D. in Social Networks and Human-Centered Computing at the Institute of Information Systems and Applications, National Tsing Hua University, Taiwan, through a collaborative program with the Institute of Information Science, Academia Sinica, Taiwan. Her research focuses on deep learning-based forgery detection, multimodal data processing, multimedia forensics, and AI-driven content authentication. She has contributed to several projects related to image and video manipulation detection, deepfake analysis, and digital forensic methodologies.



Sahibzada Adil Shahzad received his B.S. degree in Software Engineering from the University of Science and Technology Bannu, Pakistan, in 2016, and his M.S. degree in Computer Science from the Beijing Institute of Technology, China, in 2019. He is currently pursuing a Ph.D. in the Taiwan International Graduate Program in Social Networks and Human-Centered Computing, jointly offered by the Institute of Information Science, Academia Sinica, and the Department of Computer Science, National Chengchi University, Taiwan. During his bachelor's and master's studies, he worked on mobile app development and mobile augmented reality applications leveraging deep learning techniques. His current research focuses on machine learning, multimedia forensics, and audio-visual processing, with an emphasis on deepfake detection and digital media authentication.



Chia-Wen Lin (Fellow, IEEE) received his Ph.D. degree from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. Dr. Lin is currently a Distinguished Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. His research interests include image/video processing and computer vision. He has served as a Fellow Evaluation Committee member (2021–2023), BoG Members-at-Large (2022–2024), and Distinguished Lecturer (2018–2019) of the IEEE Circuits and Systems Society. He was Chair of IEEE ICME Steering Committee (2020–2021). He served as TPC Co-Chair of PCS 2022, IEEE ICIP 2019 and IEEE ICME 2010, and General Co-Chair of PCS 2024 and IEEE VCIP 2018. He received two best paper awards from VCIP 2010 and 2015. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE MULTIMEDIA. He received the Distinguished Research Award presented by National Science and Technology Council, Taiwan, in 2023, and IEEE VCIP Best Paper Award in 2015.



Yu Tsao Yu Tsao (Senior Member, IEEE) received his B.S. and M.S. degrees in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and his Ph.D. degree in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher at the National Institute of Information and Communications Technology, Tokyo, Japan, where he worked on research and product development for automatic speech recognition in multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director at the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. Additionally, he serves as a Jointly Appointed Professor in the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. Dr. Tsao's research interests include assistive oral communication technologies, audio coding, and bio-signal processing. He is an Associate Editor for both IEEE Transactions on Consumer Electronics and IEEE Signal Processing Letters. In recognition of his contributions, he received the Outstanding Research Award from the National Science and Technology Council (NSTC)—the most prestigious research honor in Taiwan—in 2023. His other accolades include the Academia Sinica Career Development Award in 2017, multiple National Innovation Awards (2018–2021 and 2023), the Future Tech Breakthrough Award in 2019, the Outstanding Elite Award from the Chung Hwa Rotary Educational Foundation (2019–2020), and the NSTC FutureTech Award in 2022. Additionally, he is the corresponding author of a paper that won the 2021 IEEE Signal Processing Society (SPS) Young Author Best Paper Award.



Hsin Min Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow. He was a Joint Professor in the Department of Computer Science and Information Engineering at National Cheng Kung University from 2014 to 2023. He was an Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing from 2016 to 2020. He currently serves a Senior Editor of APSIPA Transactions on Signal and Information Processing. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, and machine learning. He was a General Co-Chair of ISCSLP2016, ISCSLP2018, and ASRU2023 and a Technical Co-Chair of ISCSLP2010, O-COCOSDA2011, APSIPAASC2013, ISMIR2014, ASRU2019, and APSIPAASC2023. He received the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA distinguished lecturer for 2014–2015. He is a member of ISCA, ACM, and APSIPA.