IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 28 March 2023, accepted 11 April 2023, date of publication 17 April 2023, date of current version 11 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267668

RESEARCH ARTICLE

Efficient Audiovisual Fusion for Active Speaker Detection

FISEHA B. TESEMA[1], (Member, IEEE), JASON GU[2], (Senior Member, IEEE), WEI SONG[1], HONG WU[3], (Member, IEEE), SHIQIANG ZHU[1], (Member, IEEE), AND ZHEYUAN LIN[1]

[1]Interdisciplinary Innovation Research Institute, Zhejiang Lab, Zhongtai, Yuhang, Hangzhou 311121, China
[2]Electrical and Computer Engineering, Dalhousie University, Halifax, NS B3J 4R2, Canada
[3]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Corresponding author: Shiqiang Zhu (shiqiangzhu179@gmail.com)

This work was supported in part by the Zhejiang Province Ten Thousand Talents Program under Grant ID 2019R51010, in part by the Zhejiang Lab Postdoctoral start-up fund under Grant 115002-UB2107QJ, and in part by the National Natural Science Foundation of China under Grant U21A20488.

ABSTRACT Active speaker detection (ASD) refers to detecting the speaking person among visible human instances in a video. Existing methods widely employed a similar audiovisual fusion approach, the concatenation. Although such a fusion approach is often argued to help enhance performance, it must be noted that neither feature modalities play an equal role. It forces the backend network to focus on learning intramodal rather than intermodal features. Another concern is that since the concatenation doubles the fused feature dimension that feeds from the audio and video module, it creates a higher computational overhead for the backend network. To address these problems, this work hypothesizes that instead of leveraging deterministic fusion operation, employing an efficient fusion technique may assist the network in learning efficiently and improve detection accuracy. This work proposes an efficient audiovisual fusion (AVF) with fewer feature dimensions that captures the correlations between facial regions and sound signals, focusing more on the discriminative facial features and associating them with the corresponding audio features. Furthermore, previous ASD works focus only on improving ASD performance by creating a large computational overhead using complex techniques such as adding sophisticated postprocessing, applying smoothing techniques on the classifier to refine the network outputs at multiple stages, or assembling the multiple network outputs. This work proposed a simple yet effective end-to-end ASD using the newly proposed feature fusion approach, the AVF. The proposed framework attained a mAP of 84.384% on the validation set of the most challenging audiovisual speaker detection benchmark, the AVA-ActiveSpeaker. With this, this work outperformed previous works that did not apply the postprocessing tasks and attained competitive detection accuracy compared to other works that employed different postprocessing tasks. The proposed model also learns better on the unsynchronized raw AVA-ActiveSpeaker dataset. The ablation experiments under different image scale settings and noisy signals show the AFV's effectiveness and robustness than the concatenation operation.

INDEX TERMS Active speaker detection, deep learning, audiovisual fusion, human–computer interaction.

I. INTRODUCTION

Active speaker detection (ASD) aims to identify active speakers among possible candidates at a given time. It has a long history in computer vision [12]. ASD is used in various applications such as automatic video annotation [63], video conferencing [1], human-robot interactions, speech recognition, speaker diarization or re-framing video [10], speech transcription, speech enhancement [12], tracking storylines and characters in narrative content [13], [14] and facilitate the mining of training data for modeling speaker identification [15], [16]. In video conferencing, ASD allows the far-end

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su.

participants to see who is currently speaking, which is especially useful when the conference room is large, or the remote video is rendered on a small display due to small screen size, small render size, or limited bandwidth [1], [12]. In human-robot-interaction, ASD provides the robot with the knowledge of when and where a user is speaking [12]. For example, in humanoid robots/smart kiosks [5], [7], the knowledge of time and identity of the individual communicating with a robot is needed for its functioning, enabling it to manage dialogue with different speakers. However, detecting an active speaker is not explored widely and remains an open problem [17].

Existing active speaker detection works heavily rely on unimodal (visual or audio-only) or audiovisual inputs to address the problem [17], [19], [37], [41], [42], [45]. Among the unimodal-based ASD, audio-only, which provides richer information, has been widely explored in several previous works. Numerous audio-only ASD works have been explored using diarization techniques that partition the incoming audio into homogeneous streams [56], [57], [58], [59], i.e., audio sections with a single speaker. Other works interpret the audio signals that are converted into text. However, using only acoustic information performs poorly in noisy surroundings [4].

In contrast, visual-only ASD has not been widely exploited. It relies on interpreting visual information such as lip movement [43], other facial feature variations, or visual prosody information (e.g., head and lip movements) [4], [41], [42]. In [6], optical flow vectors of the mouth region are used as visual features to detect speech. However, the visual information can be easily confused by occlusions and facial movements such as facial expressions, yawning, coughing, and chewing.

Nowadays, the audiovisual approach dominates the field of ASD, for it improves the detection accuracy of ASD by increasing the robustness of the features. It leverages the inputs from both modalities by complementing a visual approach with its audio counterpart or vice-versa. References [8], [12], [38]; or uses the correlations between the audio and visual input [9], which may decrease ambiguous situations prone to cause model confusion, speaker overlap, and unrelated movement.

In the past few years, there has been a quantum leap in the performance of audiovisual-based active speaker detection due to the advances in deep learning and the recently introduced AVA-ActiveSpeaker benchmark [10]. By taking full advantage of these two aspects, several active speaker detection frameworks with appealing results have been introduced [10], [17], [18], [19], [37]. These works introduced models that employ convolutional networks to fuse local audiovisual patterns that estimate binary labels over short or long-term sequences [10], [17], [18], [19]. These models consist of frontend, fusion, and backend modules. The frontend module mainly focuses on the local motion patterns, including the frame-level, clip-level, and audio features. The fusion module fuses information from audio and visual input. The backend module focuses on the whole sequence level patterns and is designed to learn the sequence's temporal dynamics based on the features of the frontend module output.

These works have attained better results but have not addressed the main challenge in audiovisual active speaker detection: proper fusing and utilizing extracted modalities. These works [10], [17], [18], [19], [37] mainly focus on improving detection performance by introducing different frameworks that adopt or employ existing feature extraction networks combined with recurrent neural networks. None of these works exploited efficient audiovisual fusion techniques to learn; and have followed the same fusion strategy, the deterministic fusion operation, also known as the concatenation operation. Although fusing modalities with concatenation operation is often argued that it incorporates all the available feature modalities and is always beneficial to enhance the performance, it must be noted that neither feature modalities play an equal role. Thus the concatenation approach forces the backend network to focus on learning intramodal rather than intermodal features. In addition, it also creates a computational overhead to the backend network since it doubles the fused feature dimension that feeds from the audio and video module. In addition, the previous works that attained state-of-the-art detection accuracy have applied complex techniques to improve the detector's performance on the AVA-ActiveSpeaker dataset [10]. For instance, Alcázar et al. [17] used context and multi-stage refinement over a long time horizon to improve the detection performance. Chung [18] assembled the output of two networks and applied different smoothing techniques to improve the performance of the output classifier.

To better address these problems, this work hypothesizes that instead of leveraging deterministic fusion operation, employing a fusion technique with fewer dimensions may assist the network in learning efficiently and improve detection performance. Hence, this work proposes a novel audiovisual fusion (AVF) with fewer feature dimensions that captures the correlations between facial movements and sound signals, focusing more on discriminative facial features and associating them with the corresponding audio feature. This work proposes a simple, effective, end-to-end active speaker detection framework leveraging the proposed audiovisual fusion(AVF). The scope of the proposed module is to ease the computational burden from the backend network. Hence, this work did not explore the inference time of the proposed ASD model.

The proposed model comprises the frontend networks, fusion operation, and backend network. The frontend network applies two-stream architectures to process visual and audio features separately. The visual embedding module applies spatiotemporal convolution to the frame sequence and then applies the residual network (Resnet-18) to each timestamp. The audio embedding module first obtains the

Mel-frequency cepstral coefficients (MFCC) feature window from the sound signal. It then applies the most widely used audio embedding module, the modified VGGM [51], to extract audio features. The proposed fusion operation fuses the audio and visual embedding and feeds the backend network for binary classification, speaking/not speaking. The proposed framework adds independent auxiliary classification on each embedding modality with corresponding cross-entropy loss to boost the prediction network using visual and audio embedding. In summary, the key contributions of this paper are as follows:

1) This work proposes an efficient audiovisual fusion technique that learns a correlation between audio and facial movements.
2) This work proposes a novel end-to-end audiovisual-based active speaker detection framework using the proposed fusion technique.
3) This work explored the performance of different backend network sequences classifiers such as bidirectional long short-term memory (Bi-LSTM), bidirectional gated recurrent unit (Bi-GRU), and temporal convolution network (TCN) in the proposed framework. The ablation experiment showed that the Bi-GRU achieved better detection accuracy than the TCN and Bi-LSTM, and both Bi-LSTM and TCN attained almost similar detection accuracy.
4) To examine the AVF effectiveness, this work compares it with the concatenation operation on the proposed model under different experimental settings, such as different image scales and the noisy signal generated by adding a white Gaussian noise signal to the AVA-ActiveSpeaker dataset. Further, this work conducted the experiment with and without the 2 Bi-GRU layers attached after fusion on AVF-base ASD and concatenation-based ASD. These experiments reveal the effectiveness of the AVF over the concatenation approach.
5) The proposed ASD frameworks achieved a mAP of 84.384% on AVA-ActiveSpeaker validation set [10]. Compared to state-of-the-art works, the proposed framework outperformed all previous approaches that do not apply sophisticated postprocessing heuristics, assemble multiple networks, use context over a long time horizon, and refine the prediction output at different stages. The experiment also reveals that the proposed framework learns better on the raw unsynchronized AVA-ActiveSpeaker dataset.

This paper is organized as follows. Section II introduces the related works. The proposed framework is described in section III, followed by experiments in section IV. Finally, a conclusion is provided in section V.

II. RELATED WORK

Based on the input modalities used, previous active speaker detection works can be categorized into three approaches 1) audio-only, 2) visual-only, and 3) audiovisual. Here, this work gives much emphasis to the audiovisual approach.

A. AUDIO-ONLY

Audio-only active speaker detection, also known as speaker diarization, is the process of finding segments from the input audio signal associated with different speakers [45]. Bonastre et al. [61] presented a diarization approach based on binary key modeling that transforms the audio into a feature representing the speaker within the binary space. Then, employ the iterative agglomerative clustering algorithm, which forms segments of the same speaker to perform the diarization. Patino et al. [62] improved the approach using spectral clusterization. Vestman et al. [58] conduct a deep taxonomy on different features and propose a sound time-varying feature that addresses the main challenging problem in recognizing the same person regardless of the intensity of the speaker's voice. Few works [59] have exploited an end-to-end solution for active speaker detection with convolutional neural networks. References [56], [57] proposed a diarization based on Long Short-Term Memory Networks (LSTM) to capture the variations in the presenter's voice. Xie et al. [60] introduced an end-to-end utterance-level diarization framework, the new 'thinResNet' trunk architecture, which incorporates a GhostVLAD layer allowing features aggregation across time. However, in realistic situations, with far field microphones or microphone arrays, murmurs, and background noise, the task of ASD from audio is far from trivial. In addition, it cannot be associated with a visual person detection without constraining assumptions (e.g., the speaker is always visible) that do not generalize [10].

B. VISUAL-ONLY

Visual-only active speaker detection relies on visual information from the face, lip, and upper body [14]. It is vital to address the problem when audio information is unavailable or corrupted [25], [26], [27], [28]. Everingham et al. [26] assumed the motion in the lip area implied speech and used the motion of facial landmarks along the video to determine the speaker. However, this method suffers from other face/mouth motions, such as eating and yawning. Chakravarty and Tuytelaars [27] employed the dense trajectory within upper bodies to detect the active speaker. Though this approach attains nearly perfect accuracy for visible frontal faces, it is failed to generalize more complex situations, such as low resolution. Saneko et al. [28] adopted an architecture that first employs discriminative detection of visual speech and articulatory features and then performs recognition using a model that accounts for the loose synchronization of the feature streams. Reference [2] proposed an active speaker detector that uses visual prosody (lip and head movements) information before and after speech articulation to decrease the machine response time and demonstrate the discriminating power of visual prosody before and after speech articulation. Recently, Ishii et al. [3] investigated the

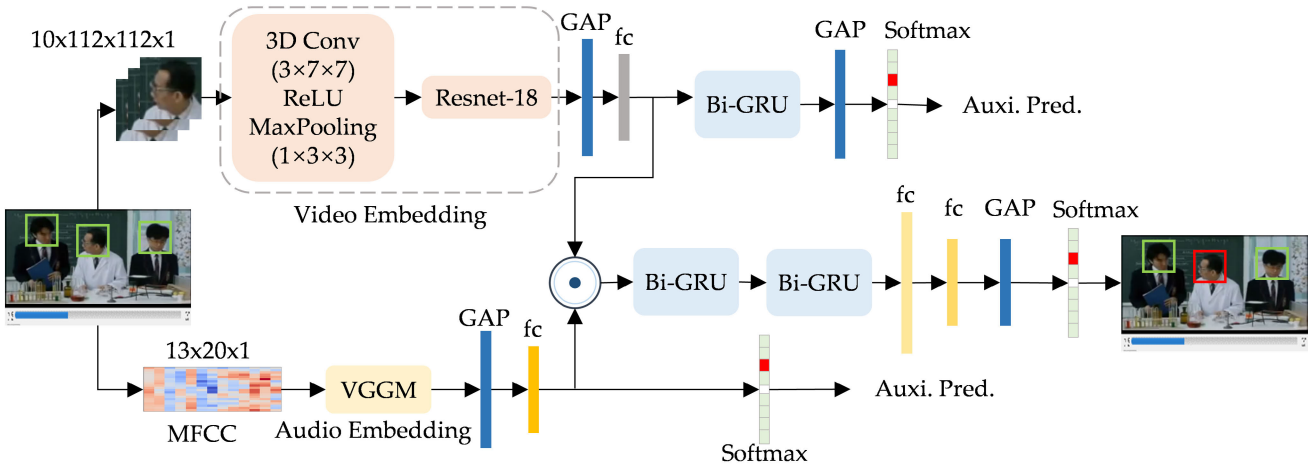


FIGURE 1. Overview of the proposed model. From the input video, the framework starts first extracts a window of T frames of cropped face regions, extract per frame audio and visual features, then feed the fused feature by audiovisual fusion (AVF) to a 2-layer bidirectional Recurrent Unit (Bi-GRU), which is followed by a 2-layer time distributed fully connected layer, 1D global average pooling, and softmax layer to obtain a final prediction. To support the main prediction network, two independent auxiliary classifications on each modality with corresponding cross-entropy loss added.

mouth-opening transition pattern (MOTP) and proposed the prediction models that predict the next-speaker and utterance interval using MOTPs.

C. AUDIOVISUAL

Multimodal learning aims at fusing information from multiple communicative modalities to create a joint representation that models the problem better than any single modalities [44]. Since audiovisual methods leverage the benefits of audiovisual inputs, it is robust to ambiguous facial movements, low-resolution images, and background noise. After the pioneering work in early 2000 by Cutler and Davis [12], audiovisual cues have paved the way for enabling ASD [12], [26], [27]. They proposed a time-delayed neural network to learn the audiovisual correlations from speech activity. Chakravarty et al. [29] re-explored leveraging audio as supervision using rich alignment between audio and visual information. Following this, [12], [16], [31] proposed a model that jointly trains an audiovisual embedding that enables more accurate active speaker detection. Friedland et al. [47] explored the application of audiovisual synchrony for active speaker localization in broadcast videos and presented an audiovisual approach for unsupervised speaker localization in meetings. Zhang et al. [48] proposed a boosting-based multimodal speaker detection algorithm for distributed meetings. An information-theoretical approach exploiting mutual correlations to associate an audio source with regions of a video stream was demonstrated by Fisher et al. [49]. In contrast, Slaney and Covell [50] proposed the audiovisual correlation used to automatically find the correct temporal synchronization between audio and a talking face. Though these previous works brought a significant breakthrough to the area, the lack of large-scale data for training and benchmarking limited these applications to in-the-wild active speaker detection in movies or consumer videos [17] for a long time. To overcome this, recently, [10] introduced AVA-ActiveSpeaker, a large-scale video dataset devised for the active speaker detection task.

A few novel approaches have been introduced following the AVA-ActiveSepaker detection dataset and its two-stream baseline network. Their proposed network learns to detect active speakers within a multi-task setting. In the AVA-ActiveSpeaker challenge of 2019, Chung [18] improved the core architecture of their previous work [12] by adding 3D convolutions and leveraging large-scale audiovisual pre-training. Reference [19] presented a framework that relied on a hybrid 3D-2D architecture, with large-scale pre-training on two multimodal datasets [12], [21]. Their method performed best when the feature embedding refines using a contrastive loss [32]. Alcázar et al. [17] proposed a context-aware model for active speaker detection that leverages cues from co-occurring speakers over long-time horizons. Huang and Koishida [37] presented an active speaker detection framework by fusing face images, dense optical flow, and audio streams. TalkNet [75] proposed a framework that uses a 3D CNN and a couple of Transformers [52], resulting in an effective large model. Another recent work, the ASD-Net [71], uses 3D-ResNet101 for encoding visual data and SincNetravanelli2018 for audio. The Unified Context Network (Unicon) [73] proposes relational context modules to capture visual (spatial) and audiovisual context based on convolutional layers. MAAS-TAN [70] proposes a different multimodal graph approach. Following MAAS-TAN, SPELL [72] presents a model that achieved superior performance by proposing an efficient graph-based framework. It is a multimodal graph from the audiovisual data and casts the active speaker detection as a graph node classification task.

SPELL differs from MAAS-TAN in several ways, where the main difference is in handling temporal context. While MAAS-TAN focuses on short-term temporal windows to construct their graphs, SPELL focuses on constructing longer-term audiovisual graphs. In MAAS-TAN, different faces are connected only between consecutive frames. In contrast, SPELL directly connects faces in a longer-term neighborhood controlled by the time threshold hyperparameter. Tesema et al. [76] proposed a simple end-to-end active two stream-based active speaker detection framework that could run in real-time, fusing visual features extracted from the image by the VGG-M and audio features extracted by MFCC.

Multimodal fusion operation has been widely exploited in problems such as speech recognition, speech enhancement [11], and action detection. Deep-irtarget [69] proposed a deep learning framework that detects targets in infrared images employing a deep neural network for feature extraction. To efficiently integrate and calibrate features, Deep-irtarget introduced a resource allocation model for features (RAF), which uses the channel and position attention blocks. Recently, graph neural network has shown an impressive ability to capture relations among support(labeled), and query(unlabeled) instances in a few-shot task [68]. However, as far as this review is concerned, none of the previous works exploited the audiovisual fusion operation for active speaker detection. Besides, the correlations between the facial regions with sound were not explored for active speaker detection. Besides, recent state-of-the-art works improved the detection accuracy with computationally expensive techniques and have employed a similar future fusion operation, which is inefficient. The proposed approach follows the baseline two-stream network but explores an orthogonal research direction by introducing an efficient fusion approach. Instead of only focusing on improving the performance of active speaker detection, this work proposes a fusion operation that reduces the computational cost of the model. Further, this work proposes a framework that can train end-to-end and perform better without additional postprocessing tasks.

III. METHOD

This section discusses the network architecture of the proposed model. As shown in Fig. 1, the network consists of two asymmetric streams for audio and video. The video stream extracts features from the raw cropped facial image inputs with the 3D-ResNet18 network. The audio stream extracts features from the MFCC features inputs with the adopted VGGM network. Each stream is followed by GAP and Fc layers to flatten the layer across the channels. The proposed audiovisual fusion module fuses each audio and visual stream and generates 128 features at each timestamp. After fusion, the two Bi-GRU layers are attached to model the joint temporal dynamic. Two time-distributed Fc and GAP layers are attached to generate the final features that feed the classifier. The output layer is a softmax layer that provides input frame labels. An independent auxiliary classifier was added to each embedding modality with corresponding cross-entropy loss to boost the prediction network using visual and audio embedding. The detailed audio and visual network will be discussed as follows.

A. AUDIO EMBEDDING

This work adopted VGGM for audio embedding, for it is successfully adopted in lip synchronization and ASD works to extract robust audio features. The adopted VGGM layer is shown in Fig. 2 (a); the network inputs a 13-dim spectrum of audio on a non-linear Mel scale of Frequency Cepstral Coefficients (MFCC) features. The features are extracted using a 25ms analysis window with a stride of 10 ms, yielding 100 audio frames every second. The input size is 20 frames in the time direction and 13 cepstral coefficients in the other direction ($13 \times 20 \times 1$) pixels. MFCC features are normalized with respect to the overall mean and variance. The proposed network was pre-trained with the improved two-stream SyncNet architecture [12], [21] for audio-to-video synchronization. The 2D global average pooling and 128-dim Fc layer are attached to an existing network to reduce the dimension of the output features.

B. VISUAL EMBEDDING

As depicted in Fig. 1, the visual embedding network inputs a video of a cropped face with a resolution of 112×112 and a frame rate of 25 fps. The frames are transformed to grayscale and normalized with respect to the overall mean and variance. The network simultaneously ingests ten stacked grayscale frames containing the visual information over the 0.4-second time frame. The visual features are extracted with the 3D-ResNet18 model first proposed in [33], which has been widely adopted for lipreading and audiovisual speech enhancement. The first set of layers applies spatiotemporal convolution to the preprocessed frame stream. The spatiotemporal convolutional layers aim to capture the face region's short-term dynamics and are helpful, even when recurrent networks are deployed for the backend [12]. It consists of a 3-dimensional (3D) convolutional layer with 64 channels kernel of $5 \times 7 \times 7$ size, followed by Batch Normalization (BN, [39]) and Rectified Linear Units (ReLU). The extracted feature maps are passed through a spatiotemporal max-pooling layer, which reduces the spatial size of the 3D feature maps. The number of parameters of the spatiotemporal frontend is $\sim$16K.

The 3D feature maps pass through within an 18-layer residual network (ResNet) (as shown in Fig. 2 (b)), one per time step. The 18-layer identity mapping version used, which was proposed for ImageNet [34]. Its building blocks consist of two convolutional layers with BN and ReLU, while the skip connections facilitate information propagation [34]. The ResNet progressively drops spatial dimensionality with max-pooling layers in the residual network.

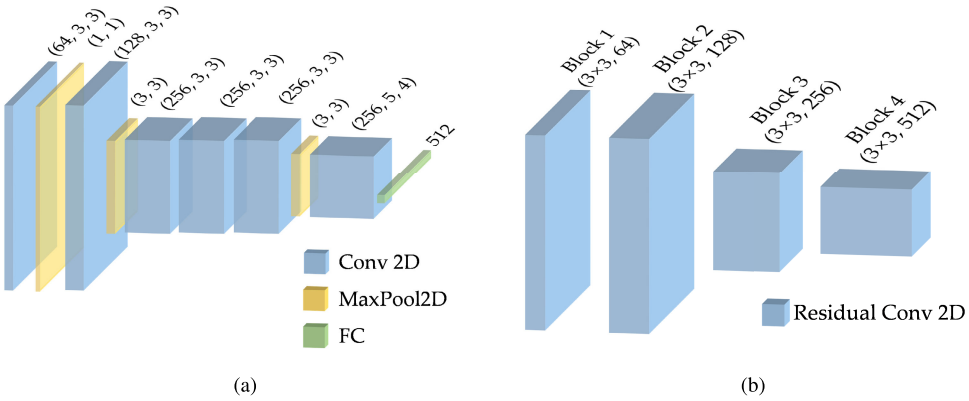


FIGURE 2. (a)The adopted VGGM (b) The adopted Resent 18.

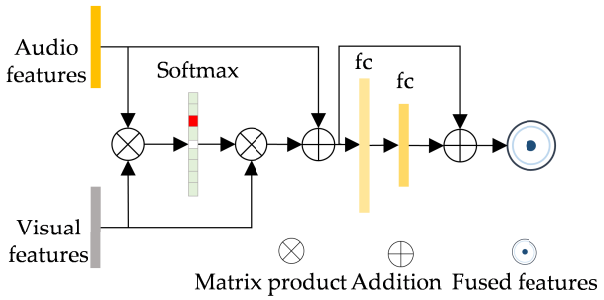


FIGURE 3. The proposed audiovisual fusion module (AVF).

The average-pooled feature is transformed into 128-dim visual embedding with the final 128-dim fully connected layer. Following the technique in [19], the subnetwork is initialized using weights pre-trained on the Lip Reading on the Wild(LRW), Lip Reading Sentences(LRS), and Multi-View Lip Reading Sentences(MV-LRS) datasets [12]. The temporal up/downsampling has not been performed for videos recorded at different frame rates during feature extraction. Because some speech/non-speech segments can be short (e.g., lasting 3-5 frames), downsampling can negatively impact frame-wise accuracy.

C. AUDIOVISUAL FUSION MODULE

This section proposes the module to handle the fusion of audiovisual modalities. Each unimodal contains different information, and the multimodal feature fusion module is to use the multimodal data's complementary information fully. Combining multiple features from feature extraction methods improves prediction accuracy. However, integrating multimodal information is a complex task for its heterogeneity. In the animal kingdom, such a task is unconsciously performed on certain brain regions, allowing beings and humans to derive multisensory conclusions. Humans can distinguish a speaking person from a not speaking person and associate the sound we hear with the corresponding visual perception from the facial or mouth region movements. Inspired by this human ability, this work proposes an audiovisual fusion module that correlates the facial region movement with audio.

Self-attention, sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence [52]. Also leveraged to capture the global and local connections [64]. It has been used successfully in various tasks, including reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations [53], [54], [55]. In this paper, the most prominent part of this work's contribution is introducing the self-attention-based module into audiovisual fusion to leverage dynamic facial cues that guide the active speaker detector. It is a self-attention [52] based cross-modal fusion module adapted to capture the correlations between face movement and the audio signals. The proposed audiovisual fusion module, henceforth abbreviated to AVF, Fig. 3 depicts a detailed illustration of the module. The fused feature F_t at each time step t is defined as follows:

$$h_t = \left(\frac{e^{(a_t . {v_t}^T)}}{\sum_{j=1}^{K} e^{((a_t . {v_t}^T)^j}}\right) v_t + a_t \tag{1}$$

$$F_t = MLP(h_t) + h_t \tag{2}$$

where $a_t \in \mathbb{R}^{(1\times 128)}$ and $v_t \in \mathbb{R}^{(1\times 128)}$, represent audio and video features at the time step t. The speaking person's facial region, specifically the mouth region [37], in a sequence of images, is sometimes related to an audio signal when the movement of the facial region along the time dimension is dynamic or not occluded. Sometimes, it does not appear when there are fewer facial dynamics, or the face is occluded. To determine how these two embeddings are related, the proposed attention first calculates the scalar product between two embedded modalities. Let us imagine embedding high facial dynamics of speaking face and speaking (load) audio since they should both encode the aspect of the speaking person, so their scalar product should be higher than if the features were completely unrelated. The scalar product between sound and visual features, which gives a relationship between the

sounds and visual embedding, has passed through a softmax activation function to make large relationships exponentially more significant. The scalar product between the softmax output and visual embedding is computed to determine the facial region's discriminative part. Finally, the new embedding (h_t) is created by combining the audio corresponding to discriminative facial regions in proportions given by the results of the scalar product of softmax function and facial region. If the sound has a strong relationship with the video, the value of audio is added to the larger part of the new embedding for the video. With this, the module captures the correlation between the modalities more easily. To improve the attention performance, the 2-layer perceptron RELU of each 256 and 128 units with residual connection was added to generate the output features F_t. The 2-layer perceptron consists of two fully connected layers of the activation function. This work proposes this kind of attention mechanism to enforce the model to focus more on the discriminative facial features and associate them with the corresponding audio features.

D. PREDICTION NETWORK

Recurrent Neural Networks (RNNs) are well-known in applications with temporal dependencies among input units, such as language modeling, machine translation, speech recognition, and image captioning. In this type of network, the hidden states represent previously seen information. Consequently, the current output depends on both the current input and the already processed outputs, and RNNs are powerful enough to maintain long-term interrelations [66]. Bidirectional RNNs (Bi-RNNs) attempt to exploit future events and previously seen information to determine the output. Since the words in a speech(sentence) are logically related to previous or subsequent words, Bi-RNNs can get forward and backward information within the sequence [65]. Similarly, Bi-RNN is commonly used in lip reading pipelines [66] and employed for a few active speaker detections [18].

A Bidirectional GRU, or BiGRU, is a sequence processing model that consists of two GRUs. One takes the input in a forward direction and the other in a backward direction [79]. It is a bidirectional recurrent neural network with only the input and forget-gates. This work employed a Bi-GRU-based backend prediction network consisting of three prediction sub-networks: the main prediction network and two auxiliary prediction networks supporting the main prediction performance. The main prediction network has 256 units of two layers of bidirectional gated recurrent units (Bi-GRU) to model the joint temporal dynamics and two layers of time distributed fully connected of size 256 and 128, respectively. Time distribution means that the same weights of a fully connected dense layer apply to each frame in the input sequence (sequence input - sequence output). Finally, the Fc's outputs are averaged in the temporal dimension with 1D global average pooling, and send the results to the softmax layer for prediction. The final layer of the main classifier outputs a probability distribution over the possible outcomes (speaking and not speaking). Likewise, the auxiliary video classifier has a 128-unit Bi-GRU layer followed by 1D global average pooling and softmax layer. The auxiliary audio classifier directly feeds the extracted features to the softmax layer.

E. IMPLEMENTATION DETAILS

The framework was implemented based on Keras [35], and the networks are trained end-to-end on a single NVIDIA Titan RTX GPU of 24GB memory. The Adam optimizer [36] is adopted, and the initial learning rate is 10^{-5}. As noted by Roth et al. [10], AVA-ActiveSpeaker has a much smaller variability than natural image datasets with a comparable size. Therefore, this work prevented overfitting during training by randomly sampling a single clip with n time contiguous crops from every face track instead of densely sampling every possible time contiguous clip of size n in the tracklet. A dropout applied, a rate of $Pdrop = 0.7$, to the time distributed fully connected layers attached after BiGRU layers.

F. LOSS FUNCTION

The loss function $l(w)$ is defined as a cross-entropy loss between the predictions and labels:

$$l(w) = -\sum_{j} y_{ij} \log(p_{ij}) + \lambda \|w\|^2, \tag{3}$$

where λ is a regularization hyperparameter, $y_i = [y_i1, \ldots, y_iN]$ is the ground truth, where $y_{ij} = 1$ if the face is Speaking and $y_{ij} = 0$ otherwise.And, prediction $p_i = [p_i1, \ldots, p_iN]$, where $p_{ij} = 1$ if the face is Speaking and $p_{ij} = 0$ otherwise

Furthermore, independent auxiliary classification networks were added on each modality with corresponding cross-entropy loss to encourage the prediction network to use audio and visual embedding. The final loss is then a combination of all terms:

$$l(w) = L_{av} + \lambda_a L_a + \lambda_v L_v, \tag{4}$$

where L_{av} is Eq. 3, L_a and L_v are the cross entropy of audio-only and visual-only networks, and following settings in [10], [37] the $\lambda_a = \lambda_b = 0.4$ were assigned to places a lower weight on the individual modality performance.

IV. EXPERIMENTS

This section discusses the experiments and evaluation results of the proposed framework on the AVA-ActiveSpeaker dataset [10]. After discussing the AVA-ActiveSpeker dataset, the following section discusses the employed evaluation metrics and acoustic noise. The remaining section discusses the ablation experiment to choose the best prediction network, the experiment to evaluate the performance of AVF and concatenation, the evaluation result on different image scale settings, the evaluation result on acoustic noise, the contribution of auxiliary supervision, the evaluation on acoustic noise, and the comparison of the proposed approach with the existing state-of-the-art approaches.

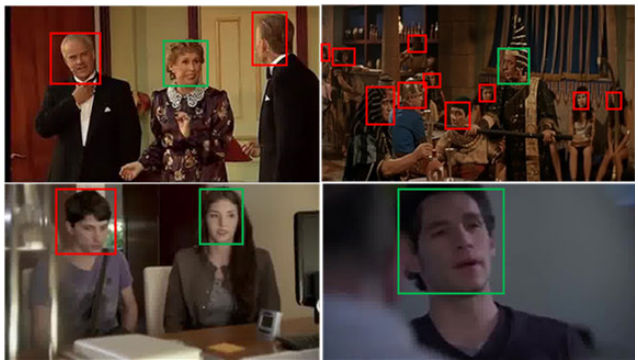


FIGURE 4. Example of labeled faces from AVA-ActiveSpeaker dataset. Red box for not speaking and a green box for speaking.

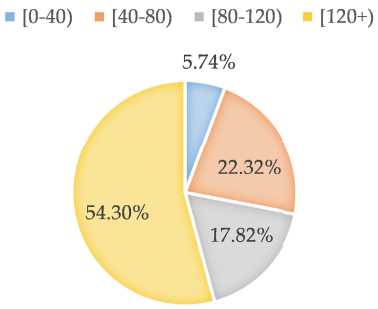


FIGURE 5. AVA-ActiveSpeaker validation set face size distribution.

A. AVA-ACTIVE SPEAKER DATASET

The AVA-ActiveSpeaker dataset [10] contains temporally labeled face tracks in video, where each face instance is labeled as speaking or not. It contains 297 movies from Youtube, with 133 for training, 33 for validation, and 131 for testing. It contains movies from film industries worldwide, leading to diverse languages, recording conditions, and speaker demographics. The dataset provides an annotation file for training and validation sets but not for the test set. Moreover, the dataset also contains many older videos in which the audio and the video appear to have been recorded separately or significantly out-of-sync [18], which makes the temporal correspondence between audio and video speech representations [12] challenging to indicate whether a person is speaking or not. A sample image is shown in Figure 4, which includes partial occlusion, different face sizes, lighting conditions, demographics, and existing activities in the dataset. The dataset consists of normalized bounding boxes for 5.3 million faces (2.6M training, 0.76M validation, and 2.0M testing) detected over 15-minute segments from each movie. In addition, the AVA-ActiveSpeaker dataset is very challenging due to the presence of low resolution (e.g., people in the distance) or occluded faces (e.g., profile faces) in the video and small labeled faces. 44.6% of labeled faces is less than 100 pixels wide, and 48.2% of the face mouth region cannot be detected by the state-of-the-art face landmark detection library Dlib [37].

B. METRICS

This work used the AVA-ActiveSpeaker dataset official evaluation technique to conduct the experiments, which computes the mean average precision (mAP) metric over the validation set.

This work also used "area under the Receiver Operating Characteristic curve (AUC)" as a metric to evaluate the performance of the proposed model under different image scales. Whenever two models are compared, a two-sided t-test is used. The P-value threshold is set to 0.05 for statistical significance.

C. ACOUSTIC NOISE

Stefanov et al. [40] evaluated their proposed voice active detection model successfully to test the effect of noise by adding stationary noise to the audio signal. Following this technique, the robustness of the proposed models against noise signals was tested by adding a stationary noise (white Gaussian Noise) to the audio signal. For every record, the noise was generated by sampling it from Gaussian distribution using standard deviation (δ), a root mean square of the signal (shown in Eq. 5) with a zero mean noisy ($\mu_{noise} = 0$). Finally, the generated noise is added to the AVA-ActiveSpeaker validation set to generate a noisy validation set.

$$\delta_{noise} = \sqrt{\frac{\sum (n_i - \mu_{noise})^2}{n}} \tag{5}$$

D. EVALUATION OF PREDICTION NETWORK

An ablation experiment has been conducted on three popular backend modules in the previous active speaker detection works the Bi-GRU, the temporal convolution network (TCN), and bidirectional long short-term memory (Bi-LSTM) to select the best backend classifier.

A TCN module comprises dilated, causal 1D convolutional layers with the same input and output lengths [77]. A Bi-LSTM is a process of making any neural network have the sequence information in both directions, backward (future to past) or forward (past to future). In bidirectional, the input flows in two directions, making a Bi-LSTM different from the regular LSTM. With the regular LSTM, input flows in one direction, either backward or forward. However, in bidirectional, input flows in both directions to preserve future and past information [78].

The Bi-GRU layer was replaced with Bi-LSTM and TCN layers to compare with the baseline. The TCN and Bi-LSTM layers have the same number of layers and units as Bi-GRU. They are evaluated under the same parameter settings (including the same dataset, same learning rate, and pre-training network). As shown in Table 1, Bi-GRU outperformed TCN and Bi-LSTM by a mAP of 1.98%. The experiment reveals no performance gap between the TCN and Bi-LSTM layer; both have attained almost the same detection performance (mAP 82.404% $\approx$ 82.359%). Hence, throughout the experiment, a Bi-GRU-based active speaker was employed.

E. EVALUATION OF FUSION OPERATIONS

The experiments have been conducted between the AVF and deterministic concatenation approach to verify the effectiveness of the proposed fusion module. These experiments were conducted on the proposed active speaker detector replacing the AVF with a concatenation operation. To further understand the effectiveness of the AVF, a comparison has been conducted against the concatenation on a new model, removing the last 2 Bi-GRU layers from the proposed ASD (without two Bi-GRU layers). For a more fair comparison, the two fusion approaches were conducted under the same settings: the dropout ratio, BNs, Bi-GRU layers, global average pooling (GAP), backbone network, optimization, learning rate, and Fc layers all are set to the same parameters. In short, the only difference is the fusion operation used. Fig. 6 illustrates the comparison of the proposed model using the proposed audiovisual fusion(AVF) and concatenation method.

1) WITH TWO Bi-GRU LAYERS

As shown in Fig. 6, the AVF outperforms the concatenation approach with small margins, mAP of 0.67%. Though the result is not statistically significant, the AVF-based ASD is more efficient than concatenation-based ASD since it feeds only 128-dim selected features to the backend network, whereas the concatenation-based ASD feeds 256-dim. With this, the AVF reduces the backend network classifier's computational burden in half compared to the concatenation operation.

2) WITHOUT TWO Bi-GRU LAYERS

Removing the two Bi-GRU layers after fusion, the AVF-based proposed ASD accuracy decreases by 0.164%mAP; on the contrary, the concatenation-based ASD drops by 10.724%mAP. From this, one can deduce that the nature of the proposed attention-based fusion module, the AVF, to handle long-range dependencies in the network [52] helps the proposed model attain better results without additional help from 2 Bi-GRU layers. The experiment reveals that the performance of concatenation-based ASD relies on the 2 Bi-GRU layers since it cannot handle long-range dependencies by itself. On the other hand, the experiment indicates that the 2 Bi-GRU layer does not help much the AVF-based ASD. Without any complications, one can use the AVF for ASD or other related applications without additional recurrent neural network layers to handle the temporal dynamics of the fused features.

On the other hand, the AVF-based ASD without 2 Bi-GRU (AVF_Wo_2biG) layers attained 82.22%mAP whereas the concatenation based without 2 Bi-GRU(concat_wo_2biG) layers attained 72.991%mAP, which is statistically significant. This work proved the hypothesis that instead of leveraging deterministic fusion operation, employing a fusion technique with fewer dimensions may assist the network in learning efficiently and improve detection performance. This work concluded that the AVF is superior and more efficient than the concatenation module.

TABLE 1. Backend network comparison.

Method	mAP%
TCN	82.404
Bi-GRU	**84.384**
Bi-LSTM	82.359

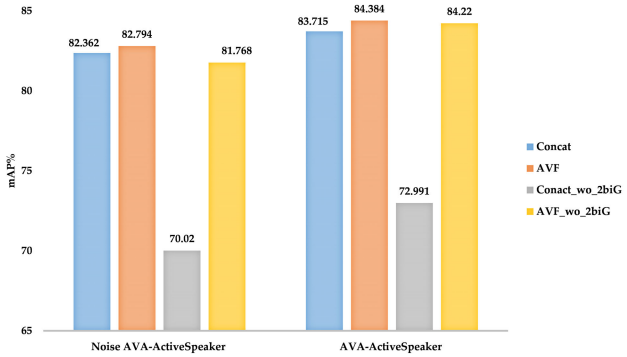


FIGURE 6. Comparison of the AVF, AVF without Bi-GRU(AVF_Wo_2biG), Concat and Concat without BiGRU(Concat_Wo_2biG) on the Noisy AVA-ActiveSpeaker, left and AVA-ActiveSpeaker dataset, right.

F. EVALUATION ON DIFFERENT IMAGE SCALE

Following the image scale settings from [40], the proposed framework was evaluated under different image scale settings on proposed ASD with and without 2 Bi-GRU layers after fusion, shown in Table 2. The AVA active speaker validation set [10] distribution is shown in Figure 5. The majority of the face size in the dataset are in the large scale (120+) range, which is 54.30%. The medium [80,120] is 17.82%, and above small [40, 80) is 22.32%. The most challenging, small-scale range [0, 40) consists of only 5.74%.

The experiment results on the proposed ASD with 2 Bi-GRU layers are not statistically significant. However, even though the AVF-based ASD inputs small-scale images, the AVF helped the proposed model select only useful features and achieved almost the same detection result as a concatenation-based model on a small scale, medium, and above small-scale settings. However, the proposed model outperformed the concatenation-based model with a mAP of 0.492% on the large-scale image, which covers most of the validation set.

On the contrary, without 2 Bi-GRU layers, the experiment result indicates statistically significant. The AVF-based ASD outperforms the concatenation-based ASD on all image scales. From this, one can infer that the AVF learns the correlation between the facial region and sounds and can pick more valuable features than the concatenation operation, which are input to the backend classification network. These experiments also revealed the robustness of the proposed fusion operation toward different image scales.

On the other hand, as shown in Table 2, the experiments on different image scales show significant performance gain as the size of the image increase on all proposed models.

TABLE 2. The performance (AUC) over different face sizes.

Method	Feature	2 Bi-GRU	[0-40)	[40-80)	[80-120)	[120+)
AVF	AV	✓	78.006	84.607	90.704	**94.171**
	V		67.446	76.832	85.715	89.411
	A		71.796	73.266	74.524	81.144
AVF	AV	✗	**78.104**	85.761	90.683	**94.085**
	V		65.972	78.172	85.849	89.393
	A		72.561	73.558	74.985	81.166
AVF_{nAux}	AV	✓	76.261	84.823	**90.874**	93.794
Concat	AV	✓	77.997	**85.195**	90.601	93.679
	V		65.770	77.845	85.941	89.266
	A		72.664	73.315	74.829	81.067
Concat	AV	✗	75.545	79.692	84.387	88.238
	V		65.25	76.00	84.732	87.928
	A		72.595	73.434	74.934	81.04

TABLE 3. Comparison with related work that employed two-stage and two-stream audiovisual framework.

Method	Contrastive loss	Postprocessing	Fusion operation & dimension	mAP%
Baseline [10]	✗	-	Concat(256)	82.100
Chung et al. (LSTM) [18]	✓	-		85.100
Chung et al. (TCN) [18]	✓	-		85.500
Chung et al. (Ensemble) [18]	✓	Assemble two models		86.100
Chung et al. (Ensemble) [18]	✓	Assemble the output of two models and apply a median filter.	Concat(512)	87.400
Chung et al. (Ensemble) [18]	✓	Assemble the output of two models and apply a Wiener filter.		87.800
Zhang et al. [19]	✓	Median filter	Concat(256)	84.000
	✗	Median filter		79.200
	✓	No smoothing		83.000
Juan et al. [17]	✓	No context	Concat(256)	79.500
		Context+No Refinements		84.400
		Context + pairwise refinement		85.200
		Context + Pairwise Refinement + MLP		85.300
		Context + Temporal		85.700
		Context + pairwise refinements + temporal refinements		87.100
Talknet [75]	✗	-	Concat(256)	92.3
Talknet [75]	✗	-	Concat(256)	82.8
Ours(without aux supervision)	✗	-	AVF(128)	83.886
Ours	✗	-	AVF(128)	84.384

TABLE 4. Comparison with SOTA that employed different approaches.

Method	Approaches	mAP%
MAAS-TAN [70]	Graph neural networks	88.8
UniCon [73]	Multiple types of contextual information	92.3
ASDNet [71]	Three stage pipeline	93.5
SPELL [72]	Graph-based representations	94.9
Ours	two-stage	84.384

The larger the input size, the better the model detects an active speaker.

G. AUXILIARY SUPERVISION HELPS?

As listed in Table 2, removing the auxiliary classifier (AVF_{nAux}), an experiment has been undertaken to observe the contribution of the auxiliary supervisions for the main classifier. The experiment reveals that the auxiliary supervision helped the model gain a mAP of 1.745% and 0.377% on small and large-scale images, respectively, and no performance gain on the medium and above small-scale images. However, as shown in Table 3 the auxiliary classifier helps the model gain a mAP of 0.498% on the overall AVA-ActiveSpekaer validation setting.

H. EVALUATION ON ACOUSTIC NOISE

The experiment has been conducted on the noisy AVA-ActiveSpeaker validation set to observe the robustness of the proposed AVF and concatenation. The noisy signals were added to the validation set according to Eq. 5. As shown in Fig. 6, the experiment result reveals that AVF-based ASD with 2 Bi-GRU layers outperformed the concatenation-based ASD with 2 Bi-GRU layers by a 0.432% mAP on the noisy AVA-ActiveSpeaker validation set and by a mAP of 0.669% on the raw AVA-ActiveSpeaker validation set. However, the result on ASD without the 2 Bi-GRU layers indicates

significant outperformance of the AVF-based ASD over the concatenation-based ASD by 11.748% mAP. The experiments proved that the proposed AVF is more robust towards a noisy signal than the deterministic operation (concatenation).

The nature of the proposed Gaussian noise is a noise that can be removed with the spectral subtraction method. However, from the experiment, this work proved that the presence of a visual stream enables the proposed models to predict the active speaker without losing much performance in the presence of competing noise. Nevertheless, as shown in Figure 6, adding Gaussian noise to the AVA-Activespeaker dataset slightly affected the performance of the proposed models.

I. COMPARISON WITH STATE-OF-THE-ART

Table 3 lists the comparison of the proposed model performance to all two-stage previous state-of-the-art works trained and evaluated on the AVA-ActiveSpeaker dataset. This work achieved a mAP of 84.384% and 83.886% without auxiliary supervision. With this result, the proposed model outperforms all two-stage previous works, except those two works [17], [18], which employed sophisticated postprocessing tasks. Computational-wise, the proposed model is less complex and efficient than two-stage previous works that achieved state-of-art detection results. For instance, Chung [18] improved ASD performance with a high computational tradeoff. They assembled the two networks' prediction output that employed LSTM and TCN classifiers. In addition, to further improve the performance, they applied a median and winder filter to remove noise from the prediction output of the classifiers in the temporal domain. Alcázar et al. [17] presented a context-based ASD that attained state-of-the-art detection results on the same dataset. However, their model consists of three phases: short-term encoder, pairwise refinement, and temporal refinement. Each phase has trained sequentially to boost the predictor's performance. Despite this, since their model was built based on the speaker's context over long time horizons, it is slow, costs higher computation, and is unsuitable for real-time applications. Without the postprocessing, their model attained less detection accuracy of 4.884% mAP than this work's and attained nearly the same result after leveraging the postprocessing. The proposed framework has three advantages over previous two-stage state-of-the-art works; 1) trained end-end-end, 2) less complex, does not employ any postprocessing tasks, and 3) it employs the proposed efficient fusion operation (the AVF) technique.

Compared to most related work of Sigtia et al. [19], the proposed framework outperforms their work that does not apply a media filtering by a mAP of 5.184%; and outperforms their model that did not employ a contrastive loss and median filtering by a mAP of 1.384% and 0.384%, respectively.

On the one hand, all the previous works listed in Table 3 used a concatenation operation, consisting of 256-dim, except Chung [18], which consists of 512-dim. Nonetheless, the proposed novel fusion operation, the AVF, has only 128 features. From this, one can infer that the proposed fusion module effectiveness. The proposed model reduced the computational burden from the backend module in half compared to most works and three-fold compared to Chung [18] works. Similarly, compared to TalkNet [75], the proposed framework attained less performance. However, computational-wise, the proposed framework uses fewer features than TalkNet. On the other hand, in similar work, the Talknet that employs 10 frames input which is the same as the proposed framework performs less by 1.58%mAP than this work.

Furthermore, all previous works, excluding the baseline work [10], enhance the detection result by employing audio-visual synchronization techniques using contrastive loss, which minimizes the distance between synchronized visual and audio embeddings and maximizes the distance for the non-synchronized pair. However, the proposed framework attained better results training end-to-end on the raw unsynchronized dataset(without the help of contrastive loss). Since the AVF helps the model join the two modalities focusing on a discriminative facial region with audio, the proposed model can learn better on unsynchronized datasets than the others. It is also simple yet effective. Since the scope of this work is to propose effective and less complex active speaker detection, further experiments adding those postprocessing tasks have not been conducted. However, one can improve the proposed model result by applying those techniques utilized in the previous works.

No further experiment has been undertaken to explore the proposed model inference time. However, since the AVF eases the computational burden of the backend network in half compared to the deterministic approach, one can easily deduce that the proposed module can improve prediction performance and ease the computation burden from the backend network leading to better inference time.

As shown in Table4, to attain a state-of-the-art result, recently, various ASD works that employ graph-based, contextual, and three-stage frameworks have been introduced. MAA-TAN [70] employed graph neural networks approach. SPELL [72] proposed a learning graph-based representation that can significantly improve the active speaker detection performance owing to its explicit spatial and temporal structure. Unicon [73] proposed a unified framework that focuses on jointly modeling multiple types of contextual information: spatial context to indicate the position and scale of each candidate's face, relational context to capture the visual relationships among the candidates and contrast audiovisual affinities with each other, and temporal context to aggregate long term information and smooth out local uncertainties. Thus, this work did not discuss the prediction and computational performance of approaches that leverage graph neural networks and contextual information for a fair comparison. ASDNet [71] is a three-stage active speaker detection pipeline consisting of the audiovisual encoding for all speakers in the clip, inter-speaker relation modeling between a reference speaker and the background speakers within each frame, and temporal modeling for the reference speaker. Compared to this work,

even though the ASDNet outperforms it, computational-wise, the ASDNet is expensive since it follows three-stage detection.

V. CONCLUSION

Previous active speaker detection followed the same fusion strategy, the concatenation, which is inefficient, and focused on improving the detection accuracy through complex postprocessing tasks. This work proposes an effective audiovisual fusion (AVF) approach that learns the correlation between facial regions and sound with less feature dimension and proposes novel efficient active speaker detection trained end-to-end. The ablation experiment on AVA-ActiveSpeaker [10] datasets under different image settings and noisy signals showed the proposed AFV's robustness and effectiveness over the deterministic fusion approach. Further, this work compared the proposed framework with state-of-the-art works, and the proposed framework outperforms all previous works that did not employ postprocessing tasks. Furthermore, the proposed framework achieved nearly the same detection accuracy as previous works that applied extensive postprocessing tasks; and learned better on the raw unsynchronized AVA-ActiveSpeaker dataset.

REFERENCES

[1] R. Cutler, R. Mehran, S. Johnson, C. Zhang, A. Kirk, O. Whyte, and A. Kowdle, "Multimodal active speaker detection and virtual cinematography for video conferencing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 4527–4531.

[2] F. Haider, S. Luz, C. Vogel, and N. Campbell, "Improving response time of active speaker detection using visual prosody information prior to articulation," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Hyderabad, India, Sep. 2018, pp. 1736–1740.

[3] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita, "Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation," *Multimodal Technol. Interact.*, vol. 3, no. 4, p. 70, Oct. 2019.

[4] F. Haider, N. Campbell, and S. Luz, "Active speaker detection in human machine multiparty dialogue using visual prosody information," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 1207–1211.

[5] A. D. Christian and B. L. Avery, "Digital smart kiosk project," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, 1998, pp. 155–162.

[6] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," in *Proc. AVSP*, 2009, pp. 151–154.

[7] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *Proc. 13th IEEE-RAS Int. Conf. Hum. Robots (Humanoids)*, Oct. 2013, pp. 203–210.

[8] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1541–1552, Dec. 2008.

[9] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME) Latest Adv. Fast Changing World Multimedia*, New York, NY, USA, Jul./Aug. 2000, pp. 1589–1592.

[10] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 4492–4496.

[11] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," 2018, *arXiv:1804.04121*.

[12] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis.*, Taiwan, Nov. 2016, pp. 251–263.

[13] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, "Movie/script: Alignment and parsing of video and text transcription," in *Proc. Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 158–171.

[14] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image Vis. Comput.*, vol. 27, no. 5, pp. 545–559, 2009.

[15] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," 2018, *arXiv:1804.03619*.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.

[17] J. L. Alcázar, F. Caba, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active speakers in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12465–12474.

[18] J. S. Chung, "Naver at ActivityNet challenge 2019—Task B active speaker detection (AVA)," 2019, *arXiv:1906.10555*.

[19] S. Sigtia, E. Marchi, S. Kajarekar, D. Naik, and J. Bridle, "Multi-task learning for speaker verification and voice trigger detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6844–6848, doi: 10.1109/ICASSP40776.2020.9054760.

[20] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" 2017, *arXiv:1705.02966*.

[21] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3965–3969.

[22] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[23] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, Edinurgh, U.K., Jul. 2005, pp. 402–414.

[24] C. Fredouille and G. Senay, "Technical improvements of the E-HMM based speaker diarization system for meeting records," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, Bethesda, MD, USA, May 2006, pp. 359–370.

[25] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, and I. Sturdy, "Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers," 2017, *arXiv:1706.00079*.

[26] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! My name is... Buffy'—Automatic naming of characters in TV video," in *Proc. BMVC*, vol. 2, no. 4, Edinburgh, U.K., Sep. 2006, pp. 1–10.

[27] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 285–301.

[28] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Beijing, China, Oct. 2005, pp. 1424–1431.

[29] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van Hamme, "Who's speaking? Audio-supervised classification of active speakers in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, Seattle, WA, USA, Nov. 2015, pp. 87–90.

[30] P. Chakravarty, J. Zegers, T. Tuytelaars, and H. Van Hamme, "Active speaker detection with audio-visual co-training," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Tokyo, Japan, Oct. 2016, pp. 312–316.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," 2018, *arXiv:1806.05622*.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1735–1742.

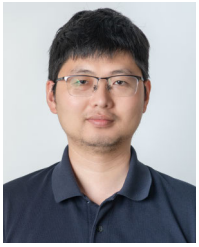
[33] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," 2017, *arXiv:1703.04105*.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 630–645.

[35] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[37] C. Huang and K. Koishida, "Improved active speaker detection based on optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 950–951.
[38] V. P. Minotto, C. R. Jung, and B. Lee, "Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1032–1044, Jun. 2014.
[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[40] K. Stefanov, J. Beskow, and G. Salvi, "Self-supervised vision-based detection of the active speaker as support for socially aware language acquisition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 12, no. 2, pp. 250–259, Jun. 2020.
[41] R. Ahmad, S. P. Raza, and H. Malik, "Visual speech detection using an unsupervised learning framework," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 2, Miami, FL, USA, Dec.2013, pp. 525–528.
[42] K. Stefanov, A. Sugimoto, and J. Beskow, "Look who's talking: Visual identification of the active speaker in multi-party human–robot interaction," in *Proc. 2nd Workshop Adv. Social Signal Process. Multimodal Interact.*, Tokyo, Japan, Nov. 2016, pp. 22–27.
[43] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 133–137, Jan. 2009.
[44] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2018.
[45] K. Stefanov, J. Beskow, and G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Proc. Grounding Lang. Understand. (GLU)*, Stockholm, Sweden, Aug. 2017.
[46] H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. Int. Conf. Image Video Retr.*, Champaign, IL, USA, Jul. 2003, pp. 488–499.
[47] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, Oct. 2009, pp. 195–202.
[48] C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola, "Boosting-based multimodal speaker detection for distributed meetings," in *Proc. IEEE Workshop Multimedia Signal Process.*, Victoria, BC, Canada, Oct. 2006, pp. 86–91.
[49] J. W. Fisher, III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2001, pp. 772–778.
[50] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, Nov. 2000, pp. 814–820.
[51] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Nov.2017, pp. 5998–6008.
[53] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," 2016, *arXiv:1606.01933*.
[54] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," 2017, *arXiv:1705.04304*.
[55] Z. Lin, M. Feng, C. Nogueira dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*.
[56] A. Sarkar, S. Dasgupta, S. K. Naskar, and S. Bandyopadhyay, "Says who? Deep learning models for joint speech recognition, segmentation and diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5229–5233.
[57] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5239–5243.
[58] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Commun.*, vol. 99, pp. 62–79, May 2018.
[59] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 4945–4949.
[60] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 5791–5795.
[61] J.-F. Bonastre, X. Anguera, G. H. Sierra, and P.-M. Bousquet, "Speaker modeling using local binary decisions," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 13–16.
[62] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD challenge," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 2813–2817.
[63] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, "Summarization of videotaped presentations: Automatic analysis of motion and gesture," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 686–696, Sep. 1998.
[64] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, "Multimodal fusion method based on self-attention mechanism," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 8843186:1–8843186:8, Sep. 2020, doi: 10.1155/2020/8843186.
[65] X. Chen, J. Du, and H. Zhang, "Lipreading with DenseNet and resBi-LSTM," *Signal, Image Video Process.*, vol. 14, no. 5, pp. 981–989, Jul. 2020, doi: 10.1007/s11760-019-01630-1.
[66] M. Oghbaie, A. Sabaghi, K. Hashemifard, and M. Akbari, "Advances and challenges in deep lip reading," 2021, *arXiv:2110.07879*.
[67] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Oct. 2019.
[68] R. Zhang, S. Yang, Q. Zhang, L. Xu, Y. He, and F. Zhang, "Graph-based few-shot learning with transformed feature propagation and optimal class allocation," *Neurocomputing*, vol. 470, pp. 247–256, Jan. 2022.
[69] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Trans. Multimedia*, vol. 24, pp. 1735–1749, 2022, doi: 10.1109/TMM.2021.3070138.
[70] J. L. Alcazar, F. C. Heilbron, A. K. Thabet, and B. Ghanem, "MAAS: Multi-modal assignation for active speaker detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 265–274.
[71] O. Kopuklu, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1193–1203.
[72] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar, "Learning long-term spatial–temporal graphs for active speaker detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 371–387.
[73] Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, and X. Chen, "UniCon: Unified context network for robust active speaker detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3964–3972, doi: 10.1145/3474085.3475275.
[74] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028.
[75] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3927–3935.
[76] F. B. B. Tesema, Z. Lin, S. Zhu, W. Song, J. Gu, and H. Wu, "End-to-end audiovisual feature fusion for active speaker detection," in *Proc. 14th Int. Conf. Digit. Image Process. (ICDIP)*, Oct. 2022, Art. no. 123422, doi: 10.1117/12.2643881.
[77] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 156–165.
[78] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 3285–3292.
[79] Z. Dai, L. Li, and W. Xu, "CFO: Conditional focused neural question answering with large-scale knowledge bases," 2016, *arXiv:1606.01994*.

FISEHA B. TESEMA (Member, IEEE) received the B.Sc. degree in computer science and IT from Haramaya University, in 2008, the M.Sc. degree in computer science from Addis Ababa University, in 2014, and the Ph.D. degree in computer science and technology from the University of Electronic Science and Technology of China, in 2020. He is currently a Postdoctoral Fellow with the Zhejiang Lab. His research interests include computer vision, human–robot interaction, multimodal learning, multimodal fusion, deep learning, and audiovisual processing.

JASON GU (Senior Member, IEEE) received the bachelor's degree in electrical engineering and information science from the University of Science and Technology of China, Hefei, China, in 1992, the master's degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1995, and the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2001. He is currently a Full Professor of electrical and computer engineering with Dalhousie University, Halifax, NS, Canada. He is also a cross-appointed Professor with the School of Biomedical Engineering, for his multidisciplinary research work. His research interests include robotics, biomedical engineering, rehabilitation engineering, neural networks, and control. He is a Fellow of the Engineering Institute of Canada.



WEI SONG received the Ph.D. degree in mechatronics engineering from the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Zhoushan, in 2013. From 2013 to 2015, he was a Postdoctoral Researcher with the Marine Research Centre, Zhejiang University. He has been an Associate Professor with Zhejiang University, since 2019. He has been a Principal Investigator with the Intelligent Robot Center, Zhejiang Lab, since 2020. His research interests include mechanical system designs, control systems, navigation systems, and artificial intelligence implemented on mobile robots.

HONG WU (Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China, in 1993, and the Ph.D. degree in pattern recognition and machine intelligence from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, in 2004. From May 2004 to May 2006, he was an Associate Researcher with the NEC Laboratories China. Since May 2006, he has been an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include computer vision and machine learning.

SHIQIANG ZHU (Member, IEEE) received the B.Eng. degree in mechanical engineering from Zhejiang University, in 1988, the M.Eng. degree in mechatronics engineering from the Beijing Institute of Technology, in 1991, and the Ph.D. degree in mechanical engineering from Zhejiang University, in 1995. He has been a Faculty Member with Zhejiang University, since 1995, where he became a Professor, in 2001. He is currently with the Ocean College, Zhejiang University, and the Director of the Zhejiang Lab. His research interests include robotics and mechatronics.

ZHEYUAN LIN received the B.Eng. degree in mechanical engineering from Xi'an Jiaotong University, in 2019, and the MSc. degree in mechanical engineering from Columbia University, in 2021. His research interests include computer vision and human–robot interaction.