A Light Weight Model for Active Speaker Detection

Junhua Liao[1], Haihan Duan[2], Kanghui Feng[1], Wanbing Zhao[1], Yanbing Yang[1,3], Liangyin Chen[1,3*]
[1] College of Computer Science, Sichuan University, Chengdu, China.
[2] The Chinese University of Hong Kong, Shenzhen, China.
[3] The Institute for Industrial Internet Research, Sichuan University, Chengdu, China.
{liaojunhua, fengkanghui, wanbingzhao}@stu.scu.edu.cn;
haihanduan@link.cuhk.edu.cn; {yangyanbing, chenliangyin}@scu.edu.cn

Abstract

Active speaker detection is a challenging task in audio-visual scenarios, with the aim to detect who is speaking in one or more speaker scenarios. This task has received considerable attention because it is crucial in many applications. Existing studies have attempted to improve the performance by inputting multiple candidate information and designing complex models. Although these methods have achieved excellent performance, their high memory and computational power consumption render their application to resource-limited scenarios difficult. Therefore, in this study, a lightweight active speaker detection architecture is constructed by reducing the number of input candidates, splitting 2D and 3D convolutions for audio-visual feature extraction, and applying gated recurrent units with low computational complexity for cross-modal modeling. Experimental results on the AVA-ActiveSpeaker dataset reveal that the proposed framework achieves competitive mAP performance (94.1% vs. 94.2%), while the resource costs are significantly lower than the state-of-the-art method, particularly in model parameters (1.0M vs. 22.5M, approximately 23×) and FLOPs (0.6G vs. 2.6G, approximately 4×). Additionally, the proposed framework also performs well on the Columbia dataset, thus demonstrating good robustness. The code and model weights are available at https://github.com/Junhua-Liao/Light-ASD.

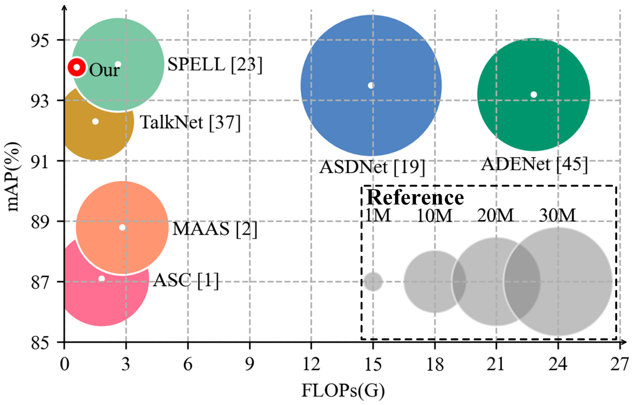


Figure 1. mAP vs. FLOPs, size ∝ parameters. This figure shows the mAP of different methods [1,2,19,23,37,45] on the benchmark and the FLOPs required to predict one frame containing three candidates. The size of the blobs is proportional to the number of model parameters. The legend shows the size of blobs corresponding to the model parameters from 1×10^6 to 30×10^6.

1. Introduction

Active speaker detection is a multi-modal task that aims to identify active speakers from a set of candidates in arbitrary videos. This task is crucial in speaker diarization [7, 42], speaker tracking [28, 29], automatic video editing [10, 20], and other applications, and thus has attracted considerable attention from both the industry and academia.

Research on active speaker detection dates back more than two decades [8, 35]. However, the lack of reliable large-scale data has delayed the development of this field. With the release of the first large-scale active speaker detection dataset, AVA-ActiveSpeaker [33], significant progress has been made in this field [15, 37, 38, 40, 47], following the rapid development of deep learning for audio-visual tasks [22]. These studies improved the performance of active speaker detection by inputting face sequences of multiple candidates simultaneously [1, 2, 47], extracting visual features with 3D convolutional neural networks [3, 19, 48], and modeling cross-modal information with complex attention modules [9, 44, 45], etc, which resulted in higher memory and computation requirements. Therefore, applying the existing methods to scenarios requiring real-time processing with limited memory and computational resources, such as automatic video editing and live television, is difficult.

[*] Corresponding author

This study proposes a lightweight end-to-end architecture designed to detect active speakers in real time, where improvements are made from the three aspects of: (a) **Single input:** inputting a single candidate face sequence with the corresponding audio; (b) **Feature extraction:** splitting the 3D convolution of visual feature extraction into 2D and 1D convolutions to extract spatial and temporal information, respectively, and splitting the 2D convolution for audio feature extraction into two 1D convolutions to extract the frequency and temporal information; (c) **Cross-modal modeling:** using gated recurrent unit (GRU) [6] with less calculation, instead of complex attention modules, for cross-modal modeling. Based on the characteristics of the lightweight architecture, a novel loss function is designed for training. Figure 1 visualizes multiple metrics of different active speaker detection approaches. The experimental results reveal that the proposed active speaker detection method (1.0M params, 0.6G FLOPs, 94.1% mAP) significantly reduces the model size and computational cost, and its performance is still comparable to that of the state-of-the-art method [23] (22.5M params, 2.6G FLOPs, 94.2% mAP) on the benchmark. Moreover, the proposed method demonstrates good robustness in cross-dataset testing. Finally, the single-frame inference time of the proposed method ranges from 0.1ms to 4.5ms, which is feasible for deployment in real-time applications.

The major contributions can be summarized as follows:

- A lightweight design is developed from the three aspects of information input, feature extraction, and cross-modal modeling; subsequently, a lightweight and effective end-to-end active speaker detection framework is proposed. In addition, a novel loss function is designed for training.

- Experiments on AVA-ActiveSpeaker [33], a benchmark dataset for active speaker detection released by Google, reveal that the proposed method is comparable to the state-of-the-art method [23], while still reducing model parameters by 95.6% and FLOPs by 76.9%.

- Ablation studies, cross-dataset testing, and qualitative analysis demonstrate the state-of-the-art performance and good robustness of the proposed method.

2. Related Work

The scientific community is increasingly interested in fusing multiple information sources to establish more effective joint representations [24]. Audio-visual learning is a common multi-modal paradigm in the video field and is used to solve tasks such as audio-visual action recognition [12, 17], audio-visual event localization [14, 32], audio-visual synchronization [4, 36], and audio-visual separation [16, 25]. The active speaker detection method studied in this study is an example of audio-visual separation.

The active speaker detection task was pioneered by Cutler and Davis [8] in the early 2000s, when they learned audio-visual correlations through time-delayed neural networks. Subsequent studies attempted to solve this task by capturing lip motion [11, 34]. Although these studies have promoted the development of this field, the lack of large-scale data for training and testing limits the application of active speaker detection in the wild. To this end, Google introduced the first large-scale video dataset, AVA-ActiveSpeaker [33], for active speaker detection, and this has resulted in the emergence of numerous novel solutions.

Alcázar *et al.* [1, 2] first exploited the temporal contextual and relational contextual information from multiple speakers to handle the active speaker detection task. Köpüklü *et al.* [19] and Min *et al.* [23] followed this idea to design structures that can better model temporal and relational contexts for improving detection performance. Subsequently, Zhang *et al.* [46, 47] introduced a spatial context to obtain a robust model by integrating three types of contextual information. Conversely, Tao *et al.* [37] achieved superior performance using cross-attention and self-attention modules to aggregate audio and visual features. Subsequently, based on this work [37], Wuerkaixi *et al.* [44] and Datta *et al.* [9] improved the performance by introducing positional encoding and improving the attention module. To better exploit the potential of the attention module, Xiong *et al.* [45] introduced multi-modal layer normalization to alleviate the distribution misalignment of audio-visual features.

In summary, existing studies focused primarily on model performance and largely ignored the cost of inputting more candidates or designing more complex models. This implies that their deployment scenarios require abundant resources, whereas the actual scenarios may not be ideal. In the field of user-generated content, TikTok and other applications provide several automatic video editing functions to assist users in their content creation. Active speaker detection provides additional possibilities for this service. As numerous users prefer creating on resource-constrained electronic devices, such as mobile phones and tablets, a lightweight model is required for deployment. In live television, active speaker detection can assist the director in cutting a shot at the current speaker, which requires the model to perform real-time detection. Therefore, a lightweight and efficient active speaker detection framework must be developed for coping with extreme environments.

3. Method

This section describes the proposed lightweight end-to-end active speaker detection approach in detail. As shown in Fig. 2, the proposed framework consists of a feature representation frontend and a speaker detection backend. The

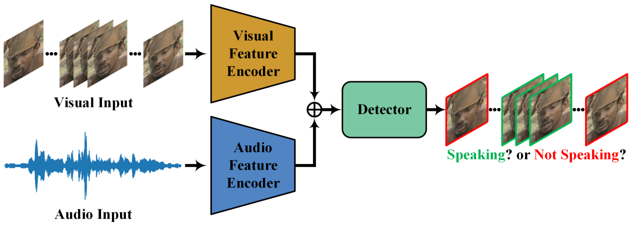


Figure 2. Overview of the proposed lightweight framework.

frontend contains visual and audio feature encoders, which encode the input candidate face sequence and the corresponding audio to obtain the features of the visual and audio signals, respectively. To fully utilize the multi-modal features, the backend detector first models the temporal context of the audio-visual features obtained by the point-wise addition of visual and audio features, and then predicts whether the current candidate is speaking.

3.1. Visual Feature Encoder

Several active speaker detection methods use 3D convolutional neural networks as visual feature encoders [3, 19, 45, 48]. Although 3D convolution can effectively extract the spatiotemporal information of face sequences, it requires numerous model parameters, and its computational cost is excessively high. To construct a lightweight visual feature encoder herein, the 3D convolution is split into 2D and 1D convolutions to extract the spatial and temporal information from the candidate face sequence, respectively. Compared with 3D convolution, this method can significantly reduce the number of model parameters and computational burden while maintaining good performance [21, 30, 39].

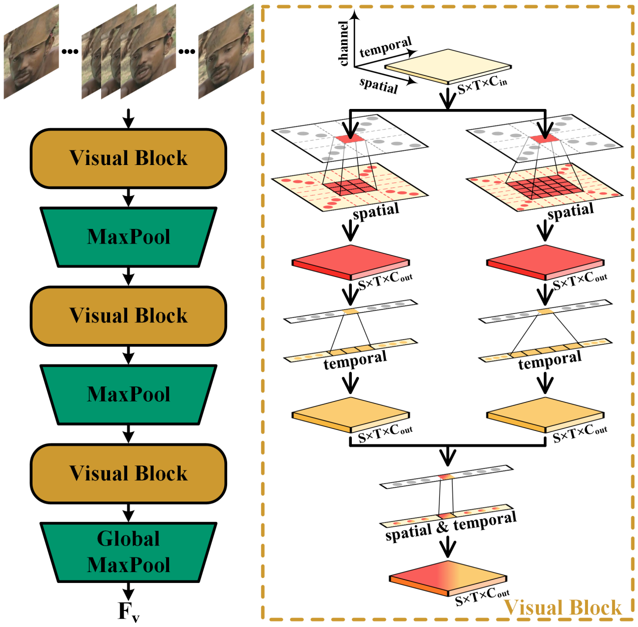


Figure 3. Architecture of visual feature encoder. The channel output dimensions C_{out} of the three visual blocks are 32, 64, and 128, respectively. The MaxPool is executed in the spatial dimension, with a kernel size of 3 and stride of 2.

The lightweight visual feature encoder is shown in Fig. 3. The encoder comprises three visual blocks, each of which contains two paths for spatiotemporal feature extraction: one is the convolution combination after 3D convolution splitting with a kernel size of 3, and the other is the convolution combination after 3D convolution splitting with a kernel size of 5. Herein, multiple paths are designed to extract features with different receptive fields and obtain abundant spatiotemporal information. Next, convolution with a kernel size of 1 integrates the features from different paths. Batch normalization and ReLU activation are performed for each convolution in the visual block. It is worth noting that all convolutions in the visual feature encoder have a stride of 1, except for the 2D convolution in the first visual block, which has a stride of 2. This design reduces the spatial dimension when extracting features, which results in smaller feature maps generated by the visual feature encoder in the subsequent feature extraction. The small-size feature map not only reduces the memory footprint, but also improves the computation speed [31]. Finally, global max pooling is performed in the spatial dimension to obtain the visual feature F_v of the candidate face sequence.

3.2. Audio Feature Encoder

Currently, Mel-frequency cepstral coefficients (MFCCs) is among the most widely used methods in audio recognition with the aim of improving the accuracy of speech activity detection [27]. Therefore, similar to most existing active speaker detection methods [9, 37, 38, 44, 45, 47], herein, a 2D feature map composed of 13D MFCCs and temporal information is extracted from the original audio signal as the input of the audio feature encoder. However, the general concept of the aforementioned studies [9, 37, 38, 44, 45, 47] that used 2D convolutional neural networks to extract audio features is not followed herein. Instead, the idea of lightweight visual blocks is adopted, and the 2D convolution is split into two 1D convolutions to extract information from the MFCCs and temporal dimensions, respectively. Figure 4 illustrates the proposed audio feature encoder architecture, comprising three audio blocks. Similar to the visual block, the audio block has two paths with different receptive fields for feature extraction, and a convolution with a kernel size of 1 is used for feature integration. It is worth mentioning that the first two max pooling layers in the audio feature encoder perform dimensionality reduction in the temporal dimension. As the original audio signal sampled by the analysis window for MFCCs typically has overlapping areas be-

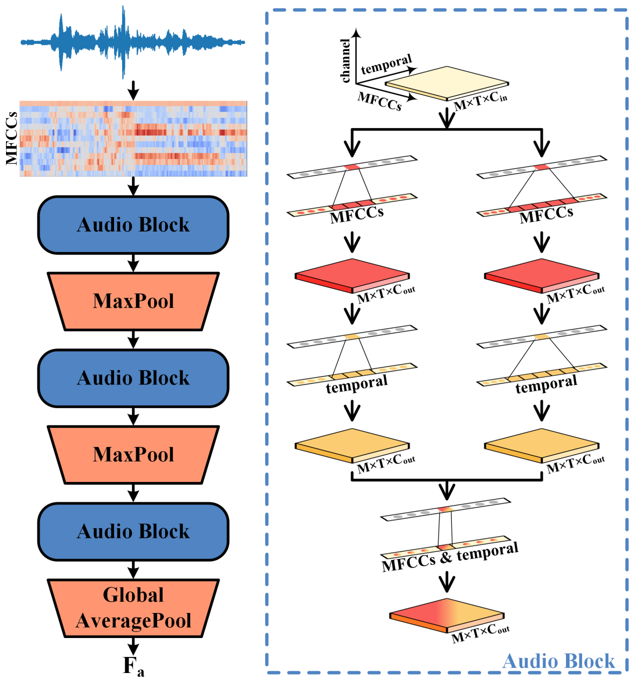


Figure 4. Architecture of the audio feature encoder. The channel output dimensions C_{out} of the three audio blocks are 32, 64, and 128, respectively. The MaxPool is executed in the temporal dimension, with a kernel size of 3 and stride of 2.

tween adjacent frames, pooling is required to maintain the temporal dimension of the audio features consistent with that of the visual features. Finally, global average pooling is performed in the MFCCs dimension to obtain the audio features F_a of the candidate.

3.3. Detector

The multi-modal features F_{av} obtained by summing the visual features F_v and audio features F_a are input into the speaker detector. The architecture of the detector is shown in Fig. 5, which is also a lightweight structure. First, the bidirectional GRU models the temporal context information of multi-modal feature F_{av}. Next, a fully connected layer (FC) predicts whether the candidate is speaking.

3.4. Loss Function

The existing active speaker detection loss function typically consists of three parts: the main classifier, the visual auxiliary classifier, and the audio auxiliary classifier [33]. Unlike previous studies [37, 44, 47], the proposed single-candidate input framework integrates visual and audio features directly without any additional cross-modal interaction. This implies that auxiliary classifiers rely only on single-modal features for prediction. In special scenarios with multiple candidates, the visual auxiliary classifier can determine whether a candidate is speaking based only on the facial information of the candidate. However, without introducing visual features, the audio auxiliary classifier can only determine whether someone is speaking, but not whether the current candidate is speaking, thus resulting in high losses. To this end, the proposed active speaker detection loss function is composed of only the main classifier and the visual auxiliary classifier.

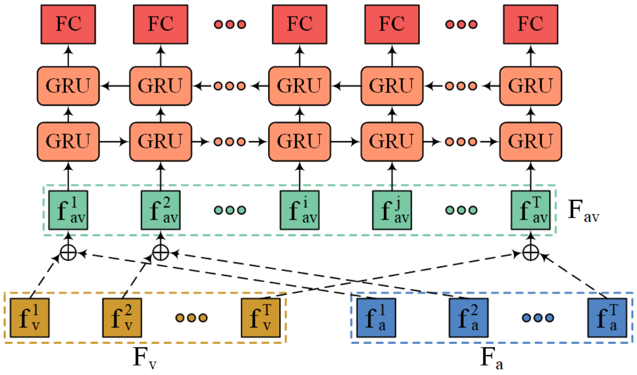


Figure 5. Architecture of the detector. f_v^i, f_a^i, and f_{av}^i represent the visual features, audio features, and multi-modal features of the i_{th} frame of the candidate sequence, respectively.

The loss function is calculated as follows:

First, the prediction result is divided by the temperature coefficient τ, and softmax is performed.

$$p_s = \frac{exp(r_{speaking}/\tau)}{exp(r_{speaking}/\tau) + exp(r_{no\_speaking}/\tau)} \tag{1}$$

where $r_{speaking}$ and $r_{no\_speaking}$ respectively represent the prediction result of whether the current candidate speaks, and p_s denotes the probability of the candidate speaking.

Note in particular that the temperature coefficient τ gradually decreases during training, as follows.

$$\tau = \tau_0 - \alpha\xi \tag{2}$$

where τ_0 is set to 1.3 as the initial temperature and α is set to 0.02 as the decay degree. ξ indicates the training epoch.

Next, the loss $\mathcal{L}$ is calculated, as follows.

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T}(g^i \log(p_s^i) + (1 - g^i)\log(1 - p_s^i)) \tag{3}$$

where p_s^i and g^i are the probability and ground truth of the candidate speaking in the i_{th} frame of the video. T refers to the number of video frames.

Finally, the complete loss function L_{asd} is obtained.

$$L_{asd} = \mathcal{L}_{av} + \lambda\mathcal{L}_v \tag{4}$$

where $\mathcal{L}_{av}$ and $\mathcal{L}_v$ denote the losses of the main classifier and the visual auxiliary classifier respectively, and λ is the weight coefficient which is set to 0.5.

Method	Single candidate?	Pre-training?	E2E?	Params(M)	FLOPs(G)	mAP(%)
ASC (CVPR'20) [1]	✗	✓	✗	23.5	1.8	87.1
MAAS (ICCV'21) [2]	✗	✓	✗	22.5	2.8	88.8
Sync-TalkNet (MLSP'22) [44]	✓	✗	✓	15.7	1.5(0.5×3)	89.8
UniCon (MM'21) [47]	✗	✓	✗	>22.4	>1.8	92.2
TalkNet (MM'21) [37]	✓	✗	✓	15.7	1.5(0.5×3)	92.3
ASD-Transformer (ICASSP'22) [9]	✓	✗	✓	>13.9	>1.5(0.5×3)	93.0
ADENet (TMM'22) [45]	✓	✗	✓	33.2	22.8(7.6×3)	93.2
ASDNet (ICCV'21) [19]	✗	✓	✗	51.3	14.9	93.5
EASEE-50 (ECCV'22) [3]	✗	✓	✓	>74.7	>65.5	94.1
SPELL (ECCV'22) [23]	✗	✓	✗	22.5	2.6	**94.2**
Our Method	✓	✗	✓	**1.0**	**0.6**(0.2×3)	94.1

Table 1. Performance comparison for methods on the validation set of the AVA-ActiveSpeaker dataset [33]. For each method, the results originate from its published paper or calculate from the open-source code. For the studies [3, 9, 47] that are not yet open source, the parameters and FLOPs of their audio-visual encoder are estimated. E2E refers to end-to-end, and FLOPs indicates the number of floating point operations required to calculate one frame containing three candidates. The FLOPs of the single candidate input method are tripled.

4. Experiment

4.1. Dataset

AVA-ActiveSpeaker. The AVA-ActiveSpeaker dataset [33] is the first large-scale standard benchmark for active speaker detection. It consists of 262 Hollywood movies, of which, 120 are training sets, 33 are validation sets, and the remaining 109 are test sets. The entire dataset contains normalized bounding boxes for 5.3 million faces, and each face detection is assigned a speaking or nonspeaking label. As a mainstream benchmark for active speaker detection tasks, this dataset contains occlusions, low-resolution faces, low-quality audio, and various lighting conditions that render it highly challenging. Note that the test set is provided for the ActivityNet challenge and is unavailable. Therefore, herein, the performance of the validation set is evaluated in a manner similar to that in previous studies [9, 38, 44, 45, 47].

Columbia. The Columbia dataset [5] is another standard test benchmark for active speaker detection. It consists of an 87-minute panel discussion video. In the video, five speakers (Bell, Boll, Lieb, Long, and Sick) take turns speaking, and 2-3 speakers are visible at any given time.

4.2. Implementation Details

Each face is reshaped into 112 × 112 pixels. The final architecture is implemented using PyTorch [26] and all experiments are performed using an NVIDIA RTX 3090 GPU (24GB). These models utilize the Adam optimizer [18] over 30 training epochs, where the learning rate is set as 0.001, with a decay rate of 0.05 for every epoch.

Evaluation metric. According to the common protocol, the metric of the AVA-ActiveSpeaker dataset is the mean Average Precision (mAP), and the metric of the Columbia dataset is the F1 score. Herein, model parameters and floating point operations (FLOPs) are reported to further measure the size and complexity of the different models.

4.3. Comparison with State-of-the-art Methods

The performance of the proposed framework is compared with that of other active speaker detection methods [1–3, 9, 19, 23, 37, 44, 45, 47] on the AVA-ActiveSpeaker validation set, and the results are summarized in Tab. 1. The four aspects of the experimental results are highlighted. (a) **Lightweight and efficient**. The mAP of the proposed method reaches 94.1%, which is only slightly inferior to the 94.2% of the state-of-the-art method SPELL [23], with 23 times fewer model parameters and 4 times fewer computations. (b) **End-to-End**. The proposed method and EASEE-50 [3] are state-of-the-art end-to-end active speaker detection methods with more than 75 times fewer model parameters and 109 times less computation. (c) **No pre-training**. Unlike approaches [1–3, 19, 23, 47] that use other large-scale datasets for pre-training models, the proposed architecture uses only the AVA-ActiveSpeaker training set to train the entire network from scratch without additional processing. (d) **Single candidate**. Existing studies [1–3, 19, 23, 47] focused on exploiting relational contextual information between speakers to improve performance. To reduce the computational burden, the proposed model inputs only a single candidate, which implies that it can make accurate predictions based on the audio and visual signals of a single candidate. In general, the results support the effectiveness and superiority of the proposed lightweight framework.

Contrary to our lightweight model design philosophy, the state-of-the-art end-to-end active speaker detection method EASEE-50 [3] uses a 3D convolutional neural network to extract the visual features of multiple input candidate face sequences. By increasing the amount of information and the complexity of the model, its performance improves to 94.1%, but the number of model parameters and FLOPs also increase to more than 74.7M and 65.5G, respectively. Multi-candidate input amplifies this disadvantage of expensive computation of the visual feature encoder based on 3D convolution, because each inference requires more computational resources to extract the visual features of multiple candidate faces. By contrast, the proposed model achieves the same mAP using only about 1% of the number of model parameters and computational cost of EASEE-50. This suggests that the small model can also achieve excellent performance in the active speaker detection task.

In addition, to evaluate the robustness of the proposed method, it is further tested on the Columbia dataset [5]. The results are presented in Tab. 2. Without fine-tuning, the proposed method achieves a state-of-the-art average F1 score of 81.1% on the Columbia dataset compared with TalkNet [37] and LoCoNet [43], showing good robustness.

Method	Speaker					
	Bell	Boll	Lieb	Long	Sick	Avg
TalkNet [37]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [43]	54.0	49.1	80.2	**80.4**	76.8	68.1
Our Method	**82.7**	**75.7**	**87.0**	74.5	**85.4**	**81.1**

Table 2. Comparison of F1-Score (%) on the Columbia dataset [5].

4.4. Ablation Studies

Kernel size. The performance of the frontend feature encoder with different convolutional kernel sizes is evaluated, and the results are presented in Tab. 3. When the encoders use convolutions with a kernel size of 3, the entire framework achieves a mAP of 93.0% with only 0.5M model parameters and 0.21G FLOPs, thus outperforming several active speaker detection methods [1, 2, 37, 44, 47]. When the size of the convolutional kernel increases from 3 to 5, the amount of information input in the feature extraction process increases, and the performance of the model is improved. However, when the convolutional kernel size is increased from 5 to 7, only the number of model parameters and the computation amount increase significantly, whereas the performance does not improve. This suggests that properly increasing the receptive field is helpful in improving the model performance. In addition, convolutions with different kernel sizes are combined and the best performance of 94.1% is achieved by combining the information under different receptive fields. This verifies the rationality and effectiveness of multipath design in visual and audio blocks.

Kernel size	Params(M)	FLOPs(G)	mAP(%)
3	0.50	0.21	93.0
5	0.77	0.42	93.4
7	1.12	0.72	93.4
3 and 5	1.02	0.63	94.1

Table 3. Impact of convolutional kernel size.

Visual feature encoder. The effectiveness of the proposed lightweight visual feature encoder is experimentally verified, and the results are presented in Tab. 4. Owing to the high computational cost of 3D convolution, numerous active speaker detection methods use 2D convolutional neural networks to extract the spatial features of face sequences and then use additional modules to extract temporal features [1, 2, 23, 37, 47]. Therefore, herein, a visual encoder from TalkNet [37] is used to verify whether the traditional ideas are more effective. This visual encoder consists of ResNet-18 [13] and a temporal module. Evidently, after the introduction of this visual encoder, the number of parameters and FLOPs of the overall architecture reaches 13.68M and 1.53G respectively, but the performance does not improve (only 92.8%). Although the large-capacity model can learn more knowledge, the input of this study is small and relatively simple face images, so a small model with an exquisite design is sufficient to complete the task of feature extraction. In addition, the dimensions of the features extracted by ResNet are relatively large, and researchers typically reduce the dimensions and conduct multi-modal modeling, which inevitably leads to information loss. Therefore, the features extracted by the proposed visual encoder are only 128 dimensions, which not only meets the design concept of lightweight, but also avoids the information loss caused by dimension reduction. Additionally, the performance of the proposed framework when the visual blocks in the visual feature encoder use 3D convolution is evaluated. Although the encoder is lightweight, 3D convolution doubles the number of model parameters and FLOPs without improving the performance. Compared with 3D convolution, the combination of 2D and 1D convolutions doubles the number of nonlinear rectifications, thereby allowing the model to represent more complex functions. Therefore, reasonably splitting 3D convolution is conducive to lightweight models and improves model performance.

Encoder	Params(M)	FLOPs(G)	mAP(%)
TalkNet [37]	13.68	1.53	92.8
3D convolution	2.06	1.56	92.9
Our Method	1.02	0.63	94.1

Table 4. Impact of visual feature encoder.

Audio feature encoder. Table 5 lists the performance of the proposed active speaker detection framework using different audio feature encoders. As an audio feature map is a 2D signal composed of MFCCs and temporal information, numerous active speaker detection methods use ResNet-18 to extract audio features [1–3, 9, 23, 47]. Therefore, first, the performance of an audio encoder based on ResNet-18 is verified. After adopting this encoder, the number of parameters in the proposed framework reaches 11.98M. However, the large-capacity model does not improve performance, probably for similar reasons to the poor performance of ResNet in the visual encoders. A large model may be prone to overfitting when information is extracted from a feature map with small dimensions. Next, the performance of using 2D convolution in audio blocks is evaluated. As the proposed audio encoder is small, the number of model parameters and FLOPs exhibits less difference before and after 2D convolution splitting. The results indicate that the performance of the audio encoder based on 2D convolution is inferior to that of the audio encoder based on 1D convolution by splitting. Perhaps the audio feature map does not have a strong spatial logic similar to that of images, so processing the MFCCs and temporal dimensions separately is more conducive to audio information aggregation.

Encoder	Params(M)	FLOPs(G)	mAP(%)
ResNet-18 [13]	11.98	0.69	93.4
2D convolution	1.12	0.63	93.6
Our Method	1.02	0.63	94.1

Table 5. Impact of audio feature encoder.

Detector. Table 6 presents the impact of the detector using different methods to process audio-visual features on model performance. Evidently, when FC is used directly for prediction without processing of audio-visual features, its mAP is only 88.0%. When a forward GRU is used for the temporal modeling of audio-visual features, the mAP increases by 4.6%. This indicates that the temporal context information of audio-visual features helps improve the performance of the active speaker detection model. However, a forward GRU can only transmit temporal information in one direction, thus causing the amount of information obtained in each frame of the sequence to be unbalanced. Therefore, herein, a bidirectional GRU is used to make each frame combine the information of the entire sequence for prediction, and achieve the best performance of 94.1%. In addition, the transformer [41] is used as an attention module to extract temporal context information of audio-visual features, and its performance is 1.1% lower than that of the forward GRU. In this attention module, all frames in the sequence have the same chance to influence the current detection frame. Although this is an effective mechanism, in this task, the information of the frames near the current detection frame is more helpful in determining whether the candidate is speaking. The forgetting mechanism of GRU renders the neighboring frames more informative, so the GRU is a better choice in this scenario.

Detector	Params(M)	FLOPs(G)	mAP(%)
None	0.82	0.63	88.0
Transformer [41]	1.02	0.63	91.5
Forward GRU	0.92	0.63	92.6
Bidirectional GRU	1.02	0.63	94.1

Table 6. Impact of the detector.

Loss function. Table 7 presents the experimental results of whether the proposed loss function L_{asd} is helpful for model training. First, when the proposed active speaker detection model is trained with the standard binary cross-entropy, it achieves 93.1% mAP on the benchmark, essentially outperforming most existing active speaker detection methods [1, 2, 9, 37, 38, 44, 47] and proving its superiority. After introducing the proposed loss function L_{asd} for training, the performance of the model improves by 1% and became 94.1%, thus indicating that the visual auxiliary classifier can better supervise the visual feature encoder. Moreover, the introduction of temperature coefficients allows the model to avoid falling into a local optimum, provides more opportunities for exploration in the early stages of training, and helps the model pay more attention to difficult samples in the later stages of training to further improve its accuracy.

Method	Params(M)	FLOPs(G)	mAP(%)
Our (without L_{asd})	1.02	0.63	93.1
Our (with L_{asd})	1.02	0.63	94.1

Table 7. Impact of the loss function.

Detection speed. The proposed framework supports dynamic length video input, so the inference time and frames per second (FPS) of the model are evaluated with different numbers of input frames. The experimental results are presented in Tab. 8. Excluding data preprocessing, the proposed framework takes 96.04ms to infer 1000 frames

Video frames	Inference time(ms)	FPS
1 (about 0.04 seconds)	4.49	223
500 (about 20 seconds)	50.28	9944
1000 (about 40 seconds)	96.04	10412

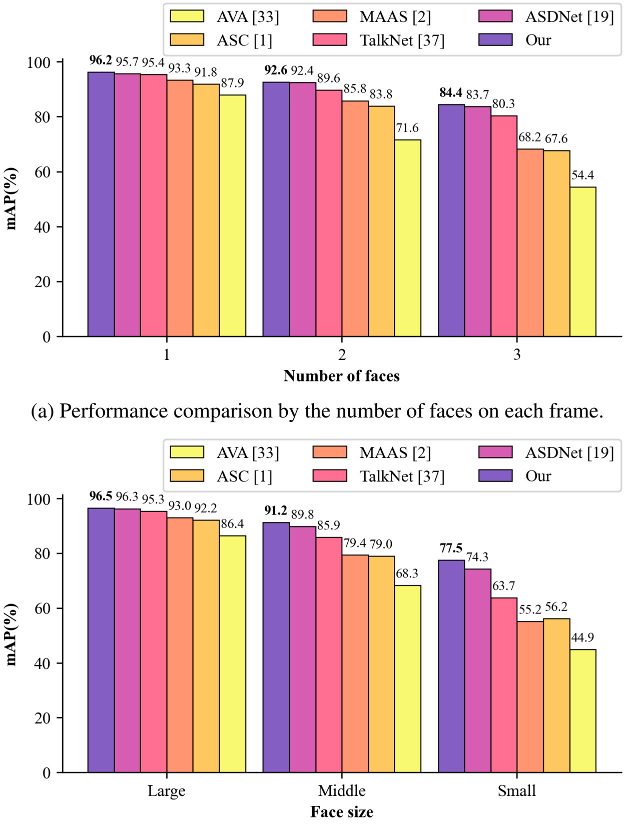
Table 8. Impact of the number of frames on the detection speed.

(about 40 seconds) of video on an NVIDIA RTX 3090 GPU, whereas the EASEE-50 [3] takes 2068.95ms for the audio-visual encoder portion alone. Even in the extreme case of a single-frame input, the inference time of the proposed framework does not exceed 4.5ms, and the FPS reaches 223, whereas the FPS of EASEE-50 is less than 36. This indicates that the proposed active speaker detection method not only meets the real-time detection requirements under different input lengths, but also has a higher detection speed compared with the state-of-the-art end-to-end method [3].

4.5. Qualitative Analysis

On the benchmark AVA-ActiveSpeaker, we break down the performance of the proposed method according to the number and size of faces, similar to the state-of-the-art methods [1,2,19,33,37]. The results are presented in Fig. 6.

First, the data are divided into three mutually exclusive groups based on the number of faces detected in a frame, which account for approximately 90% of the entire validation set. Figure 6a reports the performance of the active speaker detection methods based on the number of faces detected in the frame, and it can be seen that the performance of all methods decreases as the number of faces increases. Although the proposed method only inputs one candidate per step to reduce computational complexity, it consistently outperforms state-of-the-art methods with multiple inputs (ASC [1], MAAS [2], and ASDNet [19]) for different numbers of detected faces. The advantage of inputting multiple candidates is that the model can use not only audio-visual information but also additional relational context information to select the most likely speaker from multiple candidates. However, in the method of inputting a single candidate, the decision can be made only according to the audio-visual information of the current candidate, which has high requirements for the reliability of audio-visual features. This confirms that the proposed active speaker detection method can effectively extract and utilize audio-visual features to make accurate predictions.



(a) Performance comparison by the number of faces on each frame.

(b) Performance comparison by face size.

Figure 6. Performance breakdown. The performance of the proposed active speaker detection method and that of the previous state-of-the-art methods are evaluated on frames with one, two, and three detected faces and on faces of different sizes.

Figure 6b shows the performance of the active speaker detection methods for different face sizes. Herein, the verification set is divided into three parts according to the width of the detected faces: large (faces with widths greater than 128 pixels), middle (faces with widths between 64 and 128 pixels), and small (faces with widths less than 64 pixels). Although the performance of all the methods decreases with a decrease in face size, the advantage of the proposed method is more significant (+0.2% mAP, +1.4% mAP, +3.2% mAP). The proposed method achieves the best performance in the six scenarios subdivided by previous work, and it is the only one with mAP greater than 90% when the number of candidates is less than three or the face width is larger than 64 pixels, thus indicating that it is significantly more robust than other competing methods.

5. Conclusion

In this study, a lightweight end-to-end framework for active speaker detection is proposed. The key features of the proposed architecture include inputting a single candidate, splitting 2D and 3D convolutions for extracting audio and visual features, respectively, and using simple modules for cross-modal modeling. Experimental results on the benchmark dataset AVA-ActiveSpeaker [33] reveal that the proposed method reduces the model parameters by 95.6% and FLOPs by 76.9% compared with state-of-the-art methods, with mAP lagging by only 0.1%. In addition, the proposed active speaker detection method exhibits good robustness.

6. Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant 62072319 and 62272329); in part by the Sichuan Science and Technology Program (Grant 2023YFQ0022 and 2022YFG0041).

References

[1] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 1, 2, 5, 6, 7, 8

[2] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 1, 2, 5, 6, 7, 8

[3] Juan León Alcázar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 126–143. Springer, 2022. 1, 3, 5, 6, 7, 8

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[5] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 285–301. Springer, 2016. 5, 6

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[7] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: speaker diarisation in the wild. *arXiv preprint arXiv:2007.01216*, 2020. 1

[8] Ross Cutler and Larry Davis. Look who's talking: Speaker detection using video and audio correlation. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 3, pages 1589–1592. IEEE, 2000. 1, 2

[9] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022. 1, 2, 3, 5, 7

[10] Haihan Duan, Junhua Liao, Lehao Lin, and Wei Cai. Flad: a human-centered video content flaw detection system for meeting recordings. In *Proceedings of the 32nd Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 43–49, 2022. 1

[11] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009. 2

[12] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[14] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[15] Chong Huang and Kazuhito Koishida. Improved active speaker detection based on optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 950–951, 2020. 1

[16] Arindam Jati and Panayiotis Georgiou. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1577–1589, 2019. 2

[17] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 1, 2, 3, 5, 8

[20] Junhua Liao, Haihan Duan, Xin Li, Haoran Xu, Yanbing Yang, Wei Cai, Yanru Chen, and Liangyin Chen. Occlusion detection for automatic video editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2255–2263, 2020. 1

[21] Junhua Liao, Haihan Duan, Wanbin Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for video shot occlusion detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3154–3158. IEEE, 2022. 3

[22] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021. 1

[23] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 371–387. Springer, 2022. 1, 2, 5, 6, 7

[24] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 2

[25] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 5

[27] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019. 3

[28] Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, and Andrea Cavallaro. Audio-visual tracking of concurrent speakers. *IEEE Transactions on Multimedia*, 24:942–954, 2021. 1

[29] Xinyuan Qian, Maulik Madhavi, Zexu Pan, Jiadong Wang, and Haizhou Li. Multi-target doa estimation with an audio-visual fusion mechanism. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4280–4284. IEEE, 2021. 1

[30] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 3

[31] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 3

[32] Varshanth R Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Decompose the sounds and pixels, recompose the events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2144–2152, 2022. 2

[33] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1, 2, 4, 5, 8

[34] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell. Visual speech recognition with loosely synchronized feature streams. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1424–1431. IEEE, 2005. 2

[35] Malcolm Slaney and Michele Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. *Advances in neural information processing systems*, 13, 2000. 1

[36] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017. 2

[37] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[38] Fiseha B Tesema, Zheyuan Lin, Shiqiang Zhu, Wei Song, Jason Gu, and Hong Wu. End-to-end audiovisual feature fusion for active speaker detection. In *Fourteenth International Conference on Digital Image Processing (ICDIP 2022)*, volume 12342, pages 681–688. SPIE, 2022. 1, 3, 5, 7

[39] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[40] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1105–1114, 2021. 1

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[42] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. Speaker diarization with lstm. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 5239–5243. IEEE, 2018. 1

[43] Xizi Wang, Feng Cheng, Gedas Bertasius, and David Crandall. Loconet: Long-short context network for active speaker detection. *arXiv preprint arXiv:2301.08237*, 2023. 6

[44] Abudukelimu Wuerkaixi, You Zhang, Zhiyao Duan, and Changshui Zhang. Rethinking audio-visual synchronization for active speaker detection. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01–06. IEEE, 2022. 1, 2, 3, 4, 5, 6, 7

[45] Junwen Xiong, Yu Zhou, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Transactions on Multimedia*, pages 1–14, 2022. 1, 2, 3, 5

[46] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, and Shiguang Shan. Ictcas-ucas-tal submission to the ava-activespeaker task at activitynet challenge 2021. *The ActivityNet Large-Scale Activity Recognition Challenge*, 1(3):4, 2021. 2

[47] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972, 2021. 1, 2, 3, 4, 5, 6, 7

[48] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. *The ActivityNet Large-Scale Activity Recognition Challenge*, pages 1–4, 2019. 1, 3