International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02399-2

LR-ASD: Lightweight and Robust Network for Active Speaker Detection

Junhua Liao[1] · Haihan Duan[2,3] · Kanghui Feng[1] · Wanbing Zhao[1] · Yanbing Yang[1,4] · Liangyin Chen[1,4] · Yanru Chen[1,4]

Received: 22 June 2024 / Accepted: 17 February 2025

Abstract
Active speaker detection is a challenging task aimed at identifying who is speaking. Due to the critical importance of this task in numerous applications, it has received considerable attention. Existing studies endeavor to enhance performance at any cost by inputting information from multiple candidates and designing complex models. While these methods have achieved excellent performance, their substantial memory and computational demands pose challenges for their application to resource-limited scenarios. Therefore, in this study, a lightweight and robust network for active speaker detection, named LR-ASD, is constructed by reducing the number of input candidates, splitting 2D and 3D convolutions for audio-visual feature extraction, using a simple channel attention module for multi-modal feature fusion, and applying gated recurrent unit (GRU) with low computational complexity for temporal modeling. Results on the AVA-ActiveSpeaker dataset reveal that LR-ASD achieves competitive mean Average Precision (mAP) performance (94.5% vs. 95.2%), while the resource costs are significantly lower than the state-of-the-art method, particularly in terms of model parameters (0.84 M vs. 34.33 M, approximately 41 times) and floating point operations (FLOPs) (0.51 G vs. 4.86 G, approximately 10 times). Additionally, LR-ASD demonstrates excellent robustness by achieving state-of-the-art performance on the Talkies, Columbia, and RealVAD datasets in cross-dataset testing without fine-tuning. The project is available at https://github.com/Junhua-Liao/LR-ASD.

Keywords Active speaker detection · Lightweight · Multi-modal · Audio-visual learning

1 Introduction

Active speaker detection is a multi-modal task aimed at identifying the active speaker from a set of candidates in an arbitrary video by analyzing audio-visual information. This task serves as a crucial frontend for other downstream tasks such as speaker diarization (Qiao et al., 2024; Wang et al., 2018), speaker tracking (Qian et al., 2021; Ban et al., 2021), and automatic video editing (Liao et al., 2020, 2024), among others, thus attracting considerable attention from both industry and academia.

The research on active speaker detection dates back over two decades (Slaney & Covell, 2000; Cutler & Davis, 2000). However, the development of this field has been relatively slow due to the lack of reliable large-scale data. With the rapid advancement of deep learning for audio-visual tasks (Michel-

Communicated by Gang Hua.

✉ Yanru Chen
chenyanru@scu.edu.cn

Junhua Liao
liaojunhua@stu.scu.edu.cn

Haihan Duan
duanhaihan@smbu.edu.cn

Kanghui Feng
fengkanghui@stu.scu.edu.cn

Wanbing Zhao
wanbingzhao@stu.scu.edu.cn

Yanbing Yang
yangyanbing@scu.edu.cn

Liangyin Chen
chenliangyin@scu.edu.cn

[1] College of Computer Science, Sichuan University, Chengdu 610044, China

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China

[4] Institute for Industrial Internet Research, Sichuan University, Chengdu 610044, China

Published online: 19 March 2025

santi et al., 2021), Google released the first large-scale active speaker detection dataset, AVA-ActiveSpeaker (Roth et al., 2020), propelling the field to achieve remarkable progress (Alcázar et al., 2020; Truong et al., 2021; Tesema et al., 2022; Zhang et al., 2021b; Tao et al., 2021). These studies have significantly improved the performance of active speaker detection by incorporating multiple candidate face sequences as input (Alcázar et al., 2020, 2021; Zhang et al., 2021a), extracting visual features through 3D convolutional neural networks (Köpüklü et al., 2021; Alcázar et al., 2022; Zhang et al., 2019), and fusing visual and audio features using complex attention modules (Wuerkaixi et al., 2022; Datta et al., 2022; Xiong et al., 2022), among other approaches. However, these improvements have come at the cost of increased memory and computation requirements. Therefore, applying existing high-performance methods to real-time processing scenarios with constrained memory and computational resources, such as user-generated content creation, live television, and human-computer interactions, poses considerable difficulties.

This study proposes a lightweight and robust network for real-time end-to-end active speaker detection, named LR-ASD. LR-ASD has made lightweight improvements in the following four aspects: (a) *Single input*: inputting a single candidate face sequence with corresponding audio to minimize the memory footprint per inference; (b) *Feature extraction*: splitting 3D and 2D convolutions used for visual and audio feature extraction to separately extract spatial/frequency and temporal information reduces model parameters and computational complexity while enhancing the encoder's expressive power by doubling the number of nonlinear operations; (c) *Feature fusion*: concatenating features preserves the integrity of visual and audio information and is coupled with a simple channel attention module to address the challenge posed by the differing contributions of the two modalities to prediction; (d) *Temporal modeling*: using a module based on the computationally efficient gated recurrent unit (GRU) (Chung et al., 2014) for temporal modeling, its forget mechanism enables the current detection frame to pay more attention to the information from adjacent frames, thereby facilitating the prediction of whether the candidate is currently speaking. Figure 1 visualizes multiple metrics for different active speaker detection methods. The results demonstrate that LR-ASD (0.84 M params, 0.51 G FLOPs, 94.5% mAP) achieves a significant reduction in both model size and computational costs, while still maintaining competitive performance compared to the state-of-the-art method LoCoNet (Wang et al., 2024) (34.33 M params, 4.86 G FLOPs, 95.2% mAP) on the benchmark AVA-ActiveSpeaker dataset (Roth et al., 2020). Moreover, in cross-dataset testing without fine-tuning, LR-ASD achieves the best performance among the three datasets, Talkies (Alcázar et al., 2021), Columbia (Chakravarty & Tuytelaars, 2016), and RealVAD (Beyan et al., 2020), demonstrating remarkable robustness. Finally, the single-frame inference time of LR-ASD ranges from 0.1 to 3.9 ms, making it suitable for real-time processing.

In summary, the main contributions of this study are summarized below:

- Focusing on information input, feature extraction, feature fusion, and temporal modeling, this study proposes LR-ASD, a lightweight and robust network for active speaker detection.
- Experiments on AVA-ActiveSpeaker, a benchmark dataset for active speaker detection released by Google, reveal that LR-ASD is competitive to the state-of-the-art method LoCoNet, while still reducing model parameters by 97.6% and FLOPs by 89.5%.
- LR-ASD achieves state-of-the-art performance in cross-dataset testing without fine-tuning on the Talkies, Columbia, and RealVAD datasets, demonstrating its strong robustness. Subsequent ablation studies, quantitative analyses, and qualitative analyses further corroborate the effectiveness and robustness of LR-ASD. Finally, the good scalability of LR-ASD is validated on the EasyCom (Donley et al., 2021) and FERV39k (Wang et al., 2022) datasets.

Compared with the conference version (Liao et al., 2023), this work mainly encompasses the extension of the following aspects:

- Building upon the Light-ASD[1] in the conference version, LR-ASD improves the encoders, refines the feature fusion module, and optimizes the temporal modeling approach. Compared to Light-ASD, LR-ASD reduces model parameters and FLOPs by approximately 20%, while exhibiting performance enhancements of 0.4%, 0.2%, 5.0%, 11.2%, and 4.7% on the AVA-ActiveSpeaker, Talkies, Columbia, RealVAD, and EasyCom datasets, respectively.
- To comprehensively assess the robustness of LR-ASD, besides the AVA-ActiveSpeaker and Columbia datasets used for testing in the conference publication, we additionally introduce the Talkies (Sect. 4.4.2) and RealVAD datasets (Sect. 4.4.4) for testing, and the TalkSet dataset (Sects. 4.4.3, 4.4.4) for fine-tuning. Moreover, we introduce the EasyCom (Sect. 4.8.1) and FERV39k datasets (Sect. 4.8.2) to evaluate the scene scalability and task scalability of LR-ASD, respectively.
- We add ablation studies on data augmentation (Sect. 4.5.1), feature fusion (Sect. 4.5.5), module combination (Sect. 4.5.7), and input length (Sect. 4.5.8), along with

[1] https://github.com/Junhua-Liao/Light-ASD

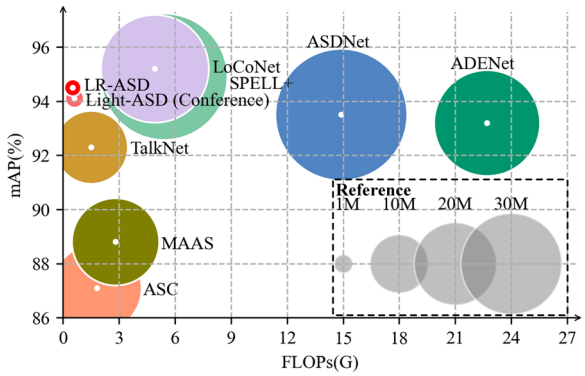


Fig. 1 mAP vs. FLOPs, size $\propto$ parameters. This figure illustrates the mAP of various methods (ASC (Alcázar et al., 2020), MAAS (Alcázar et al., 2021), TalkNet (Tao et al., 2021), ASDNet (Köpüklü et al., 2021), ADENet (Xiong et al., 2022), SPELL+ (Min et al., 2022), Light-ASD (Liao et al., 2023), and LoCoNet (Wang et al., 2024)) on the benchmark AVA-ActiveSpeaker dataset (Roth et al., 2020), along with the corresponding FLOPs necessary for predicting a single frame comprising three candidates. The size of the blobs is proportional to the number of model parameters. The legend shows the sizes of the blobs corresponding to model parameters ranging from 1×10^6 to 30×10^6

qualitative analysis (Sect. 4.7), to substantiate the effectiveness and robustness of LR-ASD.

2 Related Work

Multi-modal learning refers to the fusion of information from multiple sources to establish a more effective joint representation, thereby providing a superior means of modeling complex problems compared to isolated single-source approaches (Ngiam et al., 2011). In the video domain, audio-visual learning is a common multi-modal paradigm used to solve tasks such as audio-visual action recognition (Liu et al., 2024; Planamente et al., 2024), audio-visual synchronization (Son Chung et al., 2017; Arandjelovic & Zisserman, 2018), and audio-visual separation (Jati & Georgiou, 2019; Owens & Efros, 2018). The active speaker detection problem investigated in this study is a specific instance of audio-visual separation.

In the early 2000 s, Cutler and Davis (2000) pioneered the active speaker detection task by studying a time-delayed neural network to learn audio-visual correlations from speech activity. Subsequent research has extensively explored this field through various approaches, including capturing lip motion (Saenko et al., 2005; Matthews et al., 2002), detecting voice activity (Ramırez et al., 2004; Moattar & Homayounpour, 2009), and fusing multi-modal information (Chakravarty et al., 2016; Chung & Zisserman, 2017). However, the lack of large-scale data for both training and testing presents a substantial impediment to the widespread deployment of existing active speaker detection approaches in real-world scenarios. In 2019, Google introduced the first large-scale video dataset for active speaker detection, AVA-ActiveSpeaker (Roth et al., 2020), which marked a new epoch in the field of research pertaining to active speaker detection.

Driven by the AVA-ActiveSpeaker dataset, the dual-backbone method based on audio-visual features demonstrated considerable potential and emerged as a standard architecture for subsequent research (Chung, 2019; Zhang et al., 2019). To further improve performance, Alcázar et al. (2020) first introduced the relational contextual information from multiple speakers to handle the active speaker detection task. Following the triumph of this approach of increasing input information, subsequent studies have continuously refined and advanced this idea, resulting in significant advancements and breakthroughs (Alcázar et al., 2021; Min et al., 2022; Alcázar et al., 2022; Wang et al., 2024). Among these studies, Zhang et al. (2021a, 2021b) proposed a robust model by incorporating spatial contextual information in addition to relational and temporal contextual information. On the other hand, Tao et al. (2021) introduced cross-attention and self-attention modules to aggregate audio and visual features of single candidates, achieving excellent performance by designing complex models. Expanding on this work, Wuerkaixi et al. (2022) and Datta et al. (2022) improved the performance by introducing positional encoding and refining attention modules. To further exploit the potential of the attention module, Xiong et al. (2022) introduced a multi-modal layer normalization technique to mitigate distributional misalignment between audio and visual features.

Overall, existing research on active speaker detection has predominantly focused on improving model performance, while ignoring the costs that arise from inputting additional candidate information or designing more complex models. Deploying these methods requires abundant resources, yet resources in many real-world application scenarios are limited. For example, providing functions for user-generated content creation on mobile devices, assisting directors in real-time camera switching to the current speaker during live television, and facilitating robot interaction with speakers. In order to address the challenges posed by extreme environments, it is imperative to develop a lightweight and efficient active speaker detection framework. Therefore, this study aims to investigate the optimal trade-off between lightweight and performance, i.e., to achieve competitive performance while minimizing resource consumption.

3 Method

This section details LR-ASD, a novel lightweight and robust network for active speaker detection. As shown in Fig. 2,

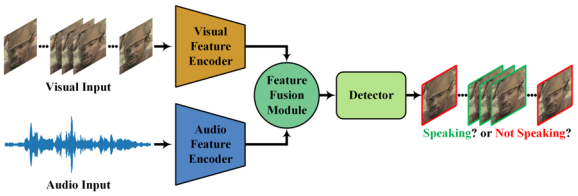


Fig. 2 The architecture of the LR-ASD

LR-ASD comprises four components. The visual and audio feature encoders process the input candidate face sequence and the corresponding audio to extract features from the visual and audio signals, respectively. The feature fusion module integrates visual and audio features into multi-modal features. The detector models the temporal context of the fused audio-visual features and subsequently predicts whether the current candidate is speaking.

3.1 Visual Feature Encoder

Due to the capability of 3D convolution to effectively extract spatiotemporal information from face sequences, some active speaker detection methods utilize it to construct the visual feature encoder (Köpüklü et al., 2021; Xiong et al., 2022; Zhang et al., 2019; Alcázar et al., 2022). Nevertheless, encoders constructed using 3D convolution not only have a large number of parameters but also entail high computational costs. Research indicates that, under careful design, decomposing 3D convolutions into 2D convolutions for spatial information extraction and 1D convolutions for temporal information extraction can achieve excellent performance and significantly reduce model parameters and computational burden (Liao et al., 2022; Duan et al., 2022, 2024).

Following this idea, we designed a lightweight visual block in the conference publication (Liao et al., 2023), as shown in Fig. 3a. It comprises two spatiotemporal feature extraction paths: 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. Finally, features from dual paths are integrated through convolution with a kernel size of 1.

Specifically, the parameter ratio R_P and FLOPs ratio R_F of these convolution combinations relative to 3D convolutions are depicted in Eqs. (1) and (2), respectively.

$$\begin{aligned} R_P &= \frac{P_{2D+1D}}{P_{3D}} \\ &= \frac{K \times K \times C_{in} \times C_{out} + K \times C_{out} \times C_{out}}{K \times K \times K \times C_{in} \times C_{out}} \\ &= \frac{K+2}{K \times K} \end{aligned} \tag{1}$$

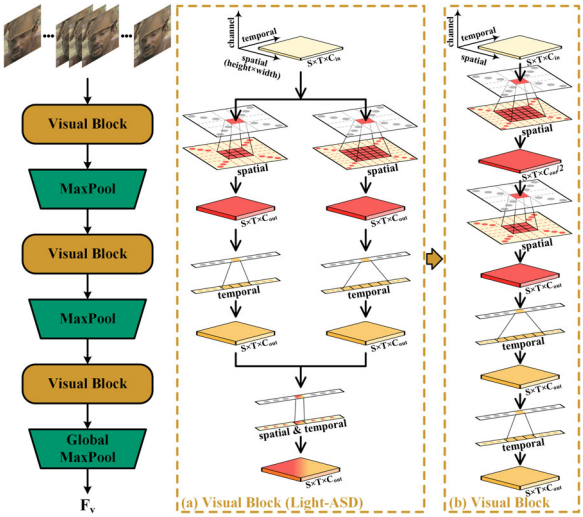


Fig. 3 Architecture of visual feature encoder. The channel output dimensions C_{out} of the three visual blocks are 32, 64, and 128, respectively. The MaxPool operation is performed along the spatial dimension with a kernel size of 3 and a stride of 2. Blocks (**a**) and (**b**) correspond to the visual blocks within the visual feature encoders of Light-ASD (Liao et al., 2023) and LR-ASD, respectively

where K denotes the kernel size of the convolution, C_{in} and C_{out} represent the dimensions of input and output feature channels, respectively, with C_{out} typically being twice C_{in}.

$$\begin{aligned} R_F &= \frac{F_{2D+1D}}{F_{3D}} = \frac{P_{2D+1D} \times H_{out} \times W_{out} \times T_{out}}{P_{3D} \times H_{out} \times W_{out} \times T_{out}} \\ &= \frac{P_{2D+1D} \times S_{out} \times T_{out}}{P_{3D} \times S_{out} \times T_{out}} = \frac{K+2}{K \times K} \end{aligned} \tag{2}$$

where K denotes the convolutional kernel size, and H_{out}, W_{out}, S_{out}, and T_{out} respectively represent the dimensions of the height, width, spatial, and temporal aspects of the output features.

As the kernel size K increases, the benefits of splitting 3D convolution in reducing parameters and FLOPs also increase. Moreover, this operation doubles the number of nonlinear rectifications, enabling the model to express more complex functions (Tran et al., 2018).

While the visual feature encoder constructed by this lightweight design exhibits high performance, the multi-branch architecture significantly increases memory access costs (Ding et al., 2021; Vasu et al., 2023). Therefore, this study designs a straight-barrel lightweight visual feature encoder, as illustrated in Fig. 3. The encoder contains three visual blocks, each consisting of four convolutional layers. Within the visual block, spatial information extraction is first performed using two layers of 2D convolutions with kernel sizes of 5 and 3, followed by temporal information extraction using two layers of 1D convolutions with kernel sizes of 5 and 3. This design further achieves model lightweight

by placing relatively complex 2D convolutions at the initial position to process features with smaller channel dimensions. It is worth noting that, in the visual feature encoder, with the exception of the initial visual block's first 2D convolution with a stride of 2, all subsequent convolutions employ a stride of 1. This design aims to reduce spatial dimensions, thereby enabling the visual feature encoder to produce smaller feature maps during subsequent feature extraction. The small-size feature maps not only reduce memory footprint but also improve computational speed (Radosavovic et al., 2020). Finally, global max pooling is performed along the spatial dimension to obtain the visual feature F_v for the candidate face sequence.

3.2 Audio Feature Encoder

In order to enhance the accuracy of speech activity detection, there are numerous methods (Ravanelli & Bengio, 2018; Krawczyk & Gerkmann, 2014; Ravanelli et al., 2019) employed for processing audio signals, with Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) being the most extensively employed (Purwins et al., 2019). Therefore, similar to most existing active speaker detection methods (Tesema et al., 2022; Wuerkaixi et al., 2022; Zhang et al., 2021b; Tao et al., 2021; Datta et al., 2022; Xiong et al., 2022; Jiang et al., 2023), this study extracts 2D feature maps consisting of 13D MFCCs and temporal information from the raw audio signals as inputs to the audio feature encoder. However, this study does not follow the conventional concept of using 2D convolutional neural networks to extract audio features as in the aforementioned studies. Instead, it adopts the idea of lightweight visual blocks, decomposing 2D convolution into two 1D convolutions to extract information from MFCCs and temporal dimensions respectively. Since the computational complexity of 2D convolution is far less than that of 3D convolution, the benefits of splitting 2D convolution in reducing model parameters and FLOPs are not significant. This operation is primarily employed to increase the number of nonlinearities within the audio block.

Figure 4 illustrates the proposed audio feature encoder, consisting of three audio blocks. Compared to the audio block in the conference publication (Liao et al., 2023), the audio block in LR-ASD has also been improved with a straight-barrel design, composed of four layers of 1D convolution. Given the overlapping frames in MFCCs analysis windows during raw audio signal sampling, dimensionality reduction is required to align the audio features with the visual features frame by frame (Tao et al., 2021; Xiong et al., 2022; Jiang et al., 2023). Therefore, the first two max pooling operations in the audio feature encoder are performed along the temporal dimension, and the final global average pooling is performed along the MFCCs dimension to obtain the audio features F_a of the candidate.

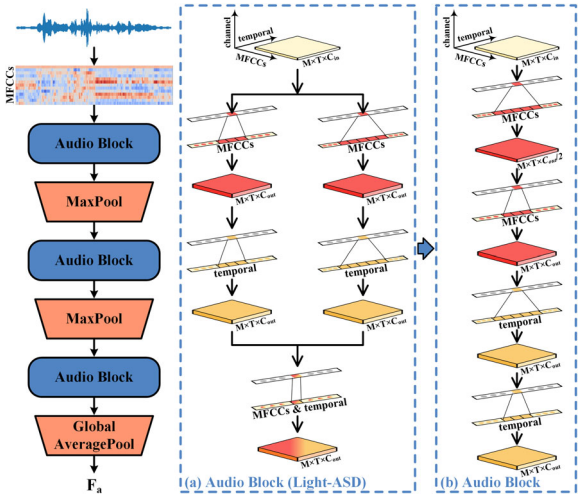


Fig. 4 Architecture of the audio feature encoder. The channel output dimensions C_{out} of the three audio blocks are 32, 64, and 128, respectively. The MaxPool operation is performed along the temporal dimension with a kernel size of 3 and a stride of 2. Blocks (**a**) and (**b**) correspond to the audio blocks within the audio feature encoders of Light-ASD (Liao et al., 2023) and LR-ASD, respectively

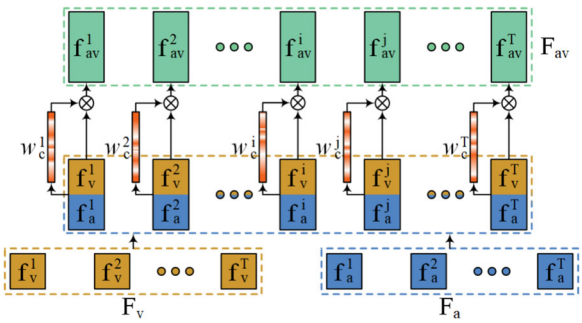


Fig. 5 Architecture of the feature fusion module. f_v^i, f_a^i, and f_{av}^i denote the visual features, audio features, and audio-visual features, respectively, of the i_{th} frame within the candidate sequence. w_c^i represents the channel weight vector of the concatenated features for the i_{th} frame in the candidate sequence, with values ranging from [0,1]

3.3 Feature Fusion Module

In Light-ASD (Liao et al., 2023), the multi-modal feature F_{av} is obtained by directly summing the visual feature F_v and the audio feature F_a. This feature fusion approach assumes equal contribution between visual and audio features. However, previous research (Köpüklü et al., 2021) and our ablation study (Sect. 4.5.7) indicate that in active speaker detection tasks, methods solely utilizing visual features outperform those solely relying on audio features by a significant margin, thus confirming the unequal contribution of visual and audio features. To better leverage both visual and audio features, this study designs a novel audio-visual feature fusion module, as illustrated in Fig. 5. This module first concatenates the visual feature F_v and the audio feature

F_a along the channel dimension to preserve the integrity of information within each modality. Subsequently, it utilizes a fully connected layer (FC) to calculate the channel weights based on the influence of information within each channel on the prediction. Finally, these weights are used to scale the concatenated features to obtain the audio-visual feature F_{av} with refined information.

3.4 Detector

Previous studies (Tao et al., 2021; Alcázar et al., 2022; Min et al., 2022; Wang et al., 2024) have shown that temporal modeling of audio-visual features can improve the performance of active speaker detection methods. The purpose of temporal modeling is to assist the model in determining whether recent lip movements match the speech activity, thereby predicting whether the candidate is speaking. Therefore, the computationally efficient GRU (Chung et al., 2014) with a forgetting mechanism is our primary choice for temporal modeling. It can model global temporal information within the sequence and filter out distant irrelevant information using the forgetting mechanism during the modeling process, enabling the current detection frame to focus more on information from adjacent frames.

Figure 6 shows the schematic of the detector's architecture. Firstly, the audio-visual feature F_{av} processed by dropout is sent to the forward and backward GRUs for temporal modeling, respectively. In order to make the detector more lightweight, the feature output dimensions are reduced to a quarter of the input dimensions during the modeling of temporal context information. Secondly, the features containing forward temporal information are concatenated with those containing backward temporal information to preserve the integrity of information from different directions. Then, the simple channel attention module within the feature fusion module is employed to calculate weights and scale the features containing bidirectional temporal information. Finally, an FC layer predicts whether the candidate is speaking.

3.5 Loss Function

The existing loss function of active speaker detection usually follows the architecture composed of the main classifier, visual auxiliary classifier, and audio auxiliary classifier (Roth et al., 2020). In special scenarios involving multiple candidates, the visual auxiliary classifier can determine whether a candidate is speaking by solely relying on the facial information of that candidate. However, in the absence of visual features, the audio auxiliary classifier can only determine the presence of speech, rather than identifying the specific candidate currently speaking, thereby resulting in high losses. To address this issue, many methods (Wuerkaixi et al., 2022; Zhang et al., 2021b; Tao et al., 2021) choose to perform

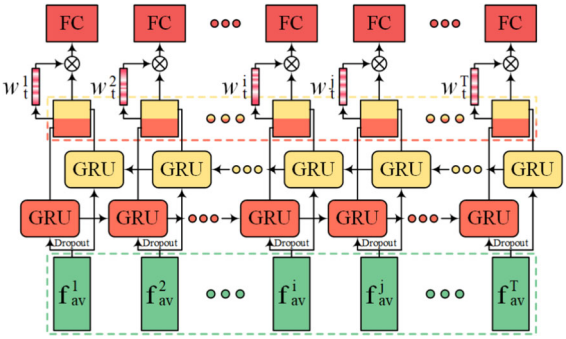


Fig. 6 Architecture of the detector. f^i_{av} represents the audio-visual features of the i_{th} frame within the candidate sequence. w^i_t represents the channel weight vector of the bidirectional temporal features for the i_{th} frame in the candidate sequence, with values ranging from [0,1]

cross-modal interactions between visual and audio features before feeding them into the auxiliary classifier. While this approach effectively alleviates the convergence difficulties of the audio auxiliary classifier, it also incurs higher computational costs, as all three classifiers are essentially making predictions based on the multi-modal audio-visual features. To mitigate computational costs, LR-ASD departs from the solution of introducing additional cross-modal interactions and instead modifies the architecture of the loss function. Specifically, the loss function employed by LR-ASD consists solely of a main classifier and a visual auxiliary classifier.

The loss function is calculated as follows:

First, apply softmax to the prediction results.

$$p_s = \frac{exp(r_{speaking})}{exp(r_{speaking}) + exp(r_{no\_speaking})} \tag{3}$$

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.

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

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T}\left(g^i \log(p_s^i) + (1-g^i)\log(1-p_s^i)\right) \tag{4}$$

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 loss function L_{asd} is obtained.

$$L_{asd} = \mathcal{L}_{av} + \lambda\mathcal{L}_v \tag{5}$$

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

Fig. 7 Example frames extracted from the datasets utilized in this work. The annotations within the green boxes denote active speakers, whereas the annotations in the red boxes represent individuals who are not speaking (Color figure online)

4 Experiment

4.1 Dataset

We primarily evaluate LR-ASD on the AVA-ActiveSpeaker dataset (Roth et al., 2020) and report its performance before and after fine-tuning on the Talkies (Alcázar et al., 2021), Columbia (Chakravarty & Tuytelaars, 2016), and RealVAD (Beyan et al., 2020) datasets. The sample frames for each dataset are shown in Fig. 7.

4.1.1 AVA-ActiveSpeaker Dataset

Google's release of the AVA-ActiveSpeaker dataset (Roth et al., 2020) marks a significant milestone as the first large-scale standard benchmark for active speaker detection. This dataset comprises 262 Hollywood movies, partitioned into three subsets: 120 for training, 33 for validation, and the remaining 109 for testing. Each movie is annotated for a duration of 15 min, including annotations for face bounding boxes, entities, and speaking labels. The complete dataset comprises normalized bounding boxes for 5.3 million faces, each of which is associated with a speaking or nonspeaking label. As a mainstream benchmark for active speaker detection evaluations, this dataset incorporates challenging factors such as occlusions, low-resolution faces, low-quality audio, and various lighting conditions, significantly increasing the difficulty level of the task. Since the test set used in the ActivityNet challenge is currently unavailable, this study evaluates the performance of the validation set in a manner similar to previous studies (Tesema et al., 2022; Xiong et al., 2022; Datta et al., 2022; Zhang et al., 2021b; Wuerkaixi et al., 2022).

4.1.2 Talkies Dataset

The Talkies dataset (Alcázar et al., 2021) is another in-the-wild active speaker detection dataset, following the AVA-ActiveSpeaker dataset. The dataset contains 23,507 face tracks extracted from 10,000 short clips, each with a duration of 1.5 s, collected from social media. Although this dataset is less extensive in scale than the AVA-ActiveSpeaker dataset, it particularly focuses on challenging scenarios featuring more speakers per frame, more diversity in terms of actors and scenes, as well as more appearances of off-screen speech.

4.1.3 Columbia Dataset

The Columbia dataset (Chakravarty & Tuytelaars, 2016) serves as another standard test benchmark for active speaker detection. This dataset annotates a 35-min segment of an 87-min panel discussion video from Columbia University, including approximate bounding boxes and active speaker labels to indicate whether the visible faces are speaking at specific time points. In the video, there are five speakers (Bell, Boll, Lieb, Long, and Sick) who take turns speaking, with two to three speakers visible at any given time.

4.1.4 RealVAD Dataset

The RealVAD dataset (Beyan et al., 2020) is similar to the Columbia dataset in that it is constructed from an 83-min panel discussion video. The video was recorded using a stationary camera that captured all the panelists, the moderator, and audiences. This dataset provides upper body bounding boxes and voice activity labels for nine panelists with different nationalities, spanning a duration of 42 min. The panelists exhibit diverse facial expressions as they move freely and engage in spontaneous actions.

4.1.5 TalkSet Dataset

The TalkSet dataset (Tao et al., 2021) is a novel hybrid dataset to address the issue of algorithmic incompatibility between the AVA-ActiveSpeaker and Columbia datasets for annotating face bounding boxes. This dataset comprises 90,000 videos with an active voice from the VoxCeleb2 dataset (Chung et al., 2018) and 60,000 videos without an active voice from the LRS3 dataset (Afouras et al., 2018). Subsequent research (Tao et al., 2021; Xiong et al., 2022; Liao et al., 2023) employing this dataset for fine-tuning resulted in a notable enhancement of the algorithm's performance on the Columbia dataset.

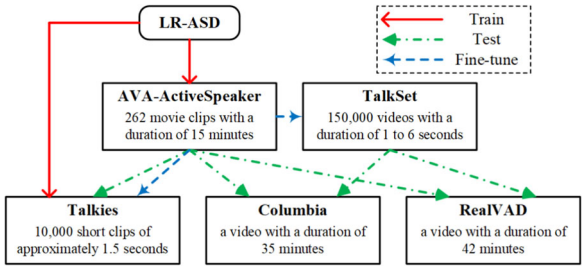


Fig. 8 The procedure for dataset utilization

4.1.6 Dataset Usage

The usage of the aforementioned datasets in this work is depicted in Fig. 8. Firstly, the LR-ASD is trained on the AVA-ActiveSpeaker dataset and its performance is evaluated on the AVA-ActiveSpeaker, Talkies, Columbia, and RealVAD datasets. Secondly, the fine-tuned and the from-scratch trained LR-ASD on the Talkies dataset are evaluated. Finally, following the precedent work (Tao et al., 2021; Xiong et al., 2022; Liao et al., 2023), LR-ASD trained from the AVA-ActiveSpeaker dataset is fine-tuned on the TalkSet dataset, and subsequently tested on the Columbia and RealVAD datasets.

4.2 Implementation Details

All facial images are standardized to a uniform size of 112×112 pixels. LR-ASD is implemented using PyTorch and subsequently trained for 40 epochs utilizing the AdamW optimizer with a weight decay of 0.01. The initial learning rate is set to 0.001, and it undergoes a decay of 0.05 per epoch. All experiments are conducted using an NVIDIA RTX 3090 GPU with 24GB of memory.

Following established protocols, the AVA-ActiveSpeaker dataset is evaluated using mean Average Precision (mAP) and Area Under Curve (AUC), the Talkies dataset evaluation employs mAP, while the Columbia and RealVAD datasets utilize the F1 score as the evaluation metric. Herein, model parameters and floating point operations (FLOPs) are reported as additional metrics for assessing the size and complexity of the various active speaker detection models.

4.3 Data Augmentation

4.3.1 Visual Data Augmentation

The augmentation techniques applied to visual data include randomly resized cropping, horizontal flipping, as well as image rotation. Specifically, the cropped region covers 49% to 100% of the standardized facial image, with the rotation angle constrained within the range of $-15°$ to $+15°$ relative to the image center. The facial images within a given candidate sequence undergo uniform processing procedures.

4.3.2 Audio Data Augmentation

To improve LR-ASD's robustness against noise, we integrate the negative sampling technique proposed by Tao et al. (2021) to augment the audio data during model training. This technique increases the number of training samples by randomly selecting an audio track from another video as noise within the same batch of the original video. This solution proves to be both straightforward and effective, as it exclusively leverages in-domain noise and interference speakers present within the training set for audio enhancement, obviating the need for external data sources beyond the training set. As a result, numerous subsequent studies (Datta et al., 2022; Wuerkaixi et al., 2022; Xiong et al., 2022; Liao et al., 2023; Wang et al., 2024) have incorporated this technique for audio augmentation.

4.4 Comparison with State-of-the-art Methods

4.4.1 Evaluation on AVA-ActiveSpeaker Dataset

Table 1 shows the performance comparison between the proposed LR-ASD and other active speaker detection methods on the AVA-ActiveSpeaker validation set. The four aspects of the experimental results are highlighted. (a) *Lightweight and efficient*. Compared to the state-of-the-art LoCoNet (Wang et al., 2024), LR-ASD reduces the model parameters by 41 times and the computations by 10 times, while mAP is only slightly lower by 0.7%, reaching 94.5%. (b) *End-to-End*. Although SPELL+ (Min et al., 2022) achieves a slightly higher mAP than LR-ASD by 0.4%, it is a two-stage method with higher complexity and computational cost compared to the end-to-end LR-ASD. (c) *No pre-training*. Unlike methods (ASC, MAAS, UniCon, ASDNet, EASEE-50, SPELL+, and LoCoNet) of employing classical neural networks as encoders and loading pre-trained weights from other large-scale datasets, LR-ASD utilizes self-designed lightweight encoders and is trained from scratch solely on the AVA-ActiveSpeaker training set. (d) *Single candidate*. Existing studies (ASC, MAAS, UniCon, ASDNet, EASEE-50, SPELL+, and LoCoNet) tend to leverage relational contextual information among speakers to improve performance. In scenarios with resource constraints, the method employing a single candidate input strategy supports longer input sequences, and the increased temporal information contributes to the enhancement of prediction accuracy. Meanwhile, results substantiate LR-ASD's capability to make accurate predictions by utilizing the audio and visual signals from individual candidates.

Table 1 Comparison of mAP (%) on the validation set of the AVA-ActiveSpeaker dataset (Roth et al., 2020)

Method	Avenue	Single candidate?	Pre-training?	End-to-End?	Params (M)	FLOPs (G)	mAP (%)	AUC (%)
ASC (Alcázar et al., 2020)	CVPR'20	✗	✔	✗	23.47	1.78	87.1	86.8
MAAS (Alcázar et al., 2021)	ICCV'21	✗	✔	✗	22.51	2.82	88.8	–
Sync-TalkNet (Wuerkaixi et al., 2022)	MLSP'22	✔	✗	✔	15.74	1.53	89.8	–
UniCon (Zhang et al., 2021b)	MM'21	✗	✔	✗	>22.35	>1.81	92.2	97.0
TalkNet (Tao et al., 2021)	MM'21	✔	✗	✔	15.74	1.53	92.3	96.8
ASD-Transformer (Datta et al., 2022)	ICASSP'22	✔	✗	✔	>13.91	>1.53	93.0	–
ADENet (Xiong et al., 2022)	TMM'22	✔	✗	✔	33.16	22.68	93.2	97.2
ASDNet (Köpüklü et al., 2021)	ICCV'21	✗	✔	✗	51.34	14.88	93.5	–
EASEE-50 (Alcázar et al., 2022)	ECCV'22	✗	✔	✔	>74.66	>65.54	94.1	–
Light-ASD (Ours) Liao et al. (2023)	CVPR'23	✔	✗	✔	1.02	0.63	94.1	97.5
SPELL (Min et al., 2022)	ECCV'22	✗	✔	✗	22.46	2.41	94.2	–
SPELL+ (Min et al., 2022)	ECCV'22	✗	✔	✗	47.32	5.37	94.9	–
LoCoNet (Wang et al., 2024)	CVPR'24	✗	✔	✔	34.33	4.86	**95.2**	**98.0**
LR-ASD	–	✔	✗	✔	**0.84**	**0.51**	94.5	97.7

Bold represents the best result in the comparison methods
For each method, the results are obtained from its published paper or calculated from the available open-source code. For studies (Zhang et al., 2021b; Datta et al., 2022; Alcázar et al., 2022) that have not yet been made open source, estimations are generated for the parameters and FLOPs of their audio-visual encoders. FLOPs stands for the number of floating point operations required to calculate a single frame containing three candidates

By increasing the amount of information and complexity of the model, LoCoNet achieves the state-of-the-art performance of 95.2%, but the number of model parameters and FLOPs also increase to 34.33 M and 4.86 G, respectively. Although LoCoNet employs a multi-candidate input strategy, it can only predict a single target candidate. Moreover, when the number of candidates is fewer than the specified count, it necessitates repeated sampling of facial sequences from the target candidate to fulfill the model's input requirements, which consumes a significant amount of computational resources. In contrast, LR-ASD achieves comparable performance using approximately 2% of the model parameters and 10% of the computational cost of LoCoNet, which indicates that the compact model is also capable of achieving exceptional performance in the active speaker detection task. Furthermore, compared to our conference publication's Light-ASD, LR-ASD achieves a 0.4% increase in mAP while simultaneously reducing model parameters by 18% and FLOPs by 19%, solidifying its position as a more competitive lightweight model.

4.4.2 Evaluation on Talkies Dataset

Results from three sets of comparative experiments conducted on the Talkies dataset are presented in Table 2. First, directly testing the models trained on the AVA-ActiveSpeaker dataset, both LR-ASD and LoCoNet achieve an optimal mAP of 88.4% on the Talkies dataset. Second, when different methods are trained from scratch on the Talkies dataset, LR-ASD achieves an mAP of 94.9%, second only to LoCoNet's 96.1%. Finally, fine-tuning various models pre-trained on the AVA-ActiveSpeaker dataset, LR-ASD achieves an mAP of 96.5% on the Talkies dataset, trailing the state-of-the-art model by a mere 0.7%. Due to the Talkies dataset's focus on challenging scenarios involving multiple candidates, the large model LoCoNet, which employs a multi-candidate input strategy, achieves state-of-the-art performance. As a lightweight solution employing a single-candidate input strategy, LR-ASD not only outperforms the conference publication's Light-ASD under different experimental settings, but also demonstrates competitiveness second only to the state-of-the-art method.

4.4.3 Evaluation on Columbia Dataset

Table 3 presents the experimental results obtained from the Columbia dataset. When the model is trained solely on the AVA-ActiveSpeaker dataset, LR-ASD achieves the highest average F1 score of 86.1% on the Columbia dataset, significantly outperforming the state-of-the-art method LoCoNet on the AVA-ActiveSpeaker dataset. Perhaps LoCoNet overfits the AVA-ActiveSpeaker dataset, resulting in its average F1 score of only 68.1% on the Columbia dataset. Compared to Light-ASD in our conference publication, LR-ASD achieves a 5% increase in average F1 score, while also ranking first in F1 score for all five speakers in this dataset. It is worth mentioning that TalkNet achieves a remarkable 40% increase in its average F1 score after undergoing fine-tuning on the

Table 2 Comparison of mAP (%) on the Talkies dataset (Alcázar et al., 2021)

Method	Training set		mAP (%)
	AVA	Talkies	
AVA Baseline (Roth et al., 2020)	✔	✗	71.5
ASC (Alcázar et al., 2020)	✔	✗	77.4
MAAS (Alcázar et al., 2021)	✔	✗	79.1
TalkNet (Tao et al., 2021)	✔	✗	86.5
EASEE-50 (Alcázar et al., 2022)	✔	✗	86.7
Light-ASD (Ours) (Liao et al., 2023)	✔	✗	88.2
LoCoNet (Wang et al., 2024)	✔	✗	**88.4**
LR-ASD	✔	✗	**88.4**
TalkNet (Tao et al., 2021)	✗	✔	93.2
EASEE-50 (Alcázar et al., 2022)	✗	✔	93.6
Light-ASD (Ours) (Liao et al., 2023)	✗	✔	94.7
LoCoNet (Wang et al., 2024)	✗	✔	**96.1**
LR-ASD	✗	✔	94.9
TalkNet (Tao et al., 2021)	✔	✔	94.4
EASEE-50 (Alcázar et al., 2022)	✔	✔	94.5
Light-ASD (Ours) (Liao et al., 2023)	✔	✔	96.2
LoCoNet (Wang et al., 2024)	✔	✔	**97.2**
LR-ASD	✔	✔	96.5

Bold represents the best result in the comparison methods

TalkSet dataset (Tao et al., 2021), ultimately achieving a state-of-the-art performance of 96.2%. To this end, we fine-tune Light-ASD and LR-ASD using the TalkSet dataset. The results indicate that LR-ASD achieves state-of-the-art performance with only approximately 5% of the parameters of TalkNet, demonstrating excellent generalization ability.

4.4.4 Evaluation on RealVAD Dataset

The comparison results of existing methods on the RealVAD dataset are summarized in Table 4. In cross-dataset testing without fine-tuning, LR-ASD achieves the highest average F1 score of 70.5%, surpassing even the existing state-of-the-art method (Beyan et al., 2020) on the RealVAD dataset by 17.5%, demonstrating its remarkable robustness. Furthermore, LR-ASD achieves the highest F1 score from eight of the nine panelists in the RealVAD dataset. Subsequently, after fine-tuning on the TalkSet dataset, the average F1 score of LR-ASD reaches 82.1%, still better than the 81.9% of Light-ASD in the conference publication. This indicates that the improvements made by LR-ASD based on Light-ASD make it have better robustness and generalization.

4.5 Ablation Studies

4.5.1 Data Augmentation

In Table 5, the impact of data augmentation techniques commonly used in active speaker detection tasks on the performance of LR-ASD is presented. Without data augmentation, LR-ASD severely overfits the training set of the AVA-ActiveSpeaker dataset, resulting in only 93.1% of its mAP on the validation set. In mitigating overfitting, augmenting either visual or audio data alone is considerably less efficacious than concurrently augmenting both visual and audio data, the latter of which elevates the mAP to 94.5%. Hence, it can be deduced that the augmentation of audio-visual data is one of the methodologies for enhancing the performance of active speaker detection methods.

4.5.2 Kernel Size

The impact of visual/audio blocks constructed using convolutions of varying kernel sizes on the LR-ASD performance is presented in Table 6. When the block is constructed with convolutions of kernel size 3, LR-ASD achieves an mAP of 93.5%, outperforming the majority of active speaker detection methods (ASC, MAAS, Sync-TalkNet, UniCon, TalkNet, ASD-Transformer, and ADENet), while utilizing merely 0.46M model parameters and 0.18G FLOPs. When the convolutional kernel size is increased from 3 to 5, bene-

Table 3 Comparison of F1-Score (%) on the Columbia dataset (Chakravarty & Tuytelaars, 2016)

Method	AVA Only	Speaker					
		Bell	Boll	Lieb	Long	Sick	Avg
TalkNet (Tao et al., 2021)	✔	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet (Wang et al., 2024)	✔	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD (Ours) (Liao et al., 2023)	✔	82.7	75.7	87.0	74.5	85.4	81.1
LR-ASD	✔	**88.8**	**77.9**	**90.3**	**85.4**	**88.3**	**86.1**
Chakravarty et al. (Chakravarty & Tuytelaars, 2016)	✗	82.9	65.8	73.6	86.9	81.8	78.2
RGB-DI (Shahid et al., 2019)	✗	86.3	93.8	92.3	76.1	86.3	87.0
Shahid et al. (Shahid et al., 2019)	✗	87.3	96.4	92.2	83.0	87.2	89.2
SyncNet (Chung & Zisserman, 2017)	✗	93.7	83.4	86.8	97.7	86.1	89.5
LWTNet (Afouras et al., 2020)	✗	92.6	82.4	88.7	94.4	95.9	90.8
RealVAD (Beyan et al., 2020)	✗	91.9	**98.9**	94.1	89.1	92.8	93.4
Truong et al. (Truong et al., 2021)	✗	95.8	88.5	91.6	96.4	97.2	93.9
S-VVAD (Shahid et al., 2021)	✗	92.4	97.2	92.3	95.5	92.5	94.0
Sharma et al. (Sharma & Narayanan, 2022)	✗	95.3	90.5	98.2	93.2	96.1	94.7
Light-ASD (Ours) (Liao et al., 2023)	✗	**97.7**	86.3	98.2	**99.0**	96.3	95.5
ADENet (Xiong et al., 2022)	✗	97.4	88.1	97.5	98.5	98.0	95.9
TalkNet (Tao et al., 2021)	✗	97.1	90.0	**99.1**	96.6	98.1	96.2
LR-ASD	✗	96.9	89.4	97.6	**99.0**	**99.2**	**96.4**

Bold represents the best result in the comparison methods

Table 4 Comparison of F1-Score (%) on the RealVAD dataset (Beyan et al., 2020)

Method	Avenue	Training set	Speaker									
			P1	P2	P3	P4	P5	P6	P7	P8	P9	Avg
TalkNet (Tao et al., 2021)	MM'21	AVA	85.8	35.5	49.8	21.5	50.7	74.6	39.6	36.9	69.0	51.5
Light-ASD (Ours) (Liao et al., 2023)	CVPR'23	AVA	75.5	**45.4**	62.4	39.4	74.1	80.4	43.8	37.6	75.1	59.3
LR-ASD	–	AVA	**88.1**	40.3	**74.0**	**73.1**	**77.5**	**82.4**	**58.1**	**57.0**	**83.9**	**70.5**
Beyan et al. (Beyan et al., 2020)	TMM'20	Columbia	53.6	51.1	41.1	50.2	37.3	50.3	56.7	53.6	69.8	51.5
S-VVAD (Shahid et al., 2021)	WACV'21	Columbia	58.3	59.3	48.0	44.8	37.3	57.4	55.6	71.3	41.1	52.6
Beyan et al. (Beyan et al., 2020)	TMM'20	RealVAD	51.6	53.5	42.9	51.7	44.4	50.5	58.7	67.9	55.8	53.0
Light-ASD (Ours) (Liao et al., 2023)	CVPR'23	AVA & TalkSet	94.4	77.3	82.5	78.2	**86.6**	85.2	**67.1**	82.3	83.8	81.9
TalkNet (Tao et al., 2021)	MM'21	AVA & TalkSet	**97.9**	**78.0**	**89.0**	**87.4**	82.8	89.9	66.3	**87.2**	**88.7**	**85.2**
LR-ASD	–	AVA & TalkSet	92.5	70.0	80.4	83.6	85.1	**91.8**	66.7	80.3	88.6	82.1

Bold represents the best result in the comparison methods
The notation 'P1-P9' represents panelists 1 to 9, respectively

fiting from the increased input information during the feature extraction process, LR-ASD's performance improved by 0.5%, reaching an mAP of 94.0%. However, blindly increasing the receptive fields does not lead to sustained performance enhancement. For example, when the size of the convolutional kernel increases from 5 to 7, LR-ASD experiences a significant increase in the number of parameters and computations, while its performance decreases. To fully leverage information gained from different receptive fields, LR-ASD explores two straight-barrel structured visual/audio blocks constructed by convolutions with kernel sizes of 3 and 5, achieving an optimal mAP of 94.5%. Among them, block (a) alternately extracts spatial/MFCCs and temporal information, while block (b) first extracts spatial/MFCCs information and then performs temporal modeling. The results indicate comparable performance between the two blocks, as the encoders constructed by each of them alternately extract spatial/MFCCs and temporal information. Moreover, the block (b) adopted by LR-ASD, based on the characteristic of having fewer feature channels in the early stage of the visual block, chooses to first use 2D convolutions with high computational complexity to extract spatial information, further reducing parameters and FLOPs compared to block (a).

Table 5 Impact of data augmentation

Method	Visual	Audio	mAP(%)
LR-ASD	✗	✗	93.1
	✔	✗	93.3
	✗	✔	93.4
	✔	✔	**94.5**

Bold represents the best result in the comparison methods

4.5.3 Visual Feature Encoder

Table 7 shows the performance of LR-ASD equipped with different visual feature encoders. Although 3D convolution is suitable for processing face sequences, numerous active speaker detection methods prefer relatively low-cost solutions. They first utilize 2D convolutional neural networks for the extraction of high-level spatial features from the face sequences, followed by temporal modeling of these features (Alcázar et al., 2020, 2021; Min et al., 2022). A representative example is the visual feature encoder used in TalkNet, which includes ResNet-18 (He et al., 2016) and a temporal module, and has been adopted in many subsequent studies (Datta et al., 2022; Wang et al., 2024; Jiang et al., 2023). To this end, we first evaluate the performance of this encoder when equipped with ResNet-18 and two lightweight networks (Howard et al., 2017; Zhang et al., 2018) respectively to extract spatial features. The results indicate that lightweight solutions can achieve comparable performance to the large-capacity model, which may be because the input in this study consists of small and relatively simple face images. Therefore, a well-designed small model is sufficient for the feature extraction task. However, the performance of these lightweight models constructed using depthwise and group convolutions in active speaker detection tasks still needs to be improved. Moreover, although depthwise convolution theoretically requires less computation, existing research (Chollet, 2017; Zhang et al., 2018) has shown that its arithmetic intensity (ratio of FLOPs to memory accesses) is too low to efficiently utilize the hardware, resulting in poor computational efficiency in practice. As for group convolution, it requires effective inter-group information interaction to compensate for the loss of global information learning ability caused by grouped processing of information (Zhang et al., 2018), while ShuffleNet's inter-group channel shuffle strategy increases memory access costs (Ding et al., 2021). Finally, the feature dimensions extracted by these models are relatively large, and these studies typically perform dimensionality reduction before conducting multi-modal modeling, which inevitably leads to information loss.

Differing from the classic idea, the visual feature encoders of Light-ASD and LR-ASD achieve lightweight design by splitting the 3D convolutions, enabling temporal modeling to be conducted during the process of spatial feature extraction. Meanwhile, our encoders extract visual features with only 128 dimensions, avoiding the information loss caused by dimension reduction. The experiments indicate that compared to the visual feature encoder of Light-ASD, the encoder of LR-ASD is more lightweight and performs superiorly. Furthermore, LR-ASD's visual feature encoder adopts a straight-barrel structure, resulting in lower memory access costs (Ding et al., 2021; Vasu et al., 2023) compared to the

Table 6 Impact of convolutional kernel size

Kernel size	Params (M)	FLOPs (G)	mAP (%)
3	**0.46**	**0.18**	93.5
5	0.73	0.42	94.0
7	1.08	0.72	93.9
(a) 3 and 5 (S/M_5, T_5, S/M_3, T_3)	0.86	0.57	94.2
(b) 3 and 5 (S/M_5, S/M_3, T_5, T_3)	0.84	0.51	**94.5**

Bold represents the best result in the comparison methods
S, M, and T represent spatial, MFCCs, and temporal dimensions, respectively, with subscript numbers indicating kernel size

Table 7 Impact of visual feature encoder

Encoder	Params (M)	FLOPs (G)	mAP (%)
TalkNet (Tao et al., 2021)	13.64	1.53	92.1
ShuffleNet (Zhang et al., 2018)	6.97	**0.45**	92.1
MobileNet (Howard et al., 2017)	6.19	0.48	91.8
LR-ASD (3D Convolution)	1.37	1.02	93.6
Light-ASD (Ours)	0.98	0.63	94.3
LR-ASD	**0.84**	0.51	**94.5**

Bold represents the best result in the comparison methods

Table 8 Impact of audio feature encoder

Encoder	Params (M)	FLOPs (G)	mAP (%)
ResNet-18 (He et al., 2016)	11.84	0.57	94.1
LR-ASD (2D Convolution)	1.06	**0.51**	94.0
Light-ASD (Ours)	0.88	**0.51**	94.4
LR-ASD	**0.84**	**0.51**	**94.5**

Bold represents the best result in the comparison methods

multi-branch structure of Light-ASD. Finally, we also evaluate the performance of LR-ASD's visual feature encoder without splitting, i.e., using 3D convolutions. Although the encoder is lightweight, employing 3D convolution doubles FLOPs while reducing performance. Compared to 3D convolutions, the combination of 2D and 1D convolutions doubles the number of nonlinear rectifications, enabling the model to represent more complex functions. Therefore, reasonably splitting 3D convolution is not only beneficial for model lightweightness but also improves performance.

4.5.4 Audio Feature Encoder

The impact of different audio feature encoders on the performance of LR-ASD is shown in Table 8. Many active speaker detection methods utilize ResNet-18 as an audio feature encoder to process 2D audio feature maps consisting of MFCCs and temporal information (Datta et al., 2022; Alcázar et al., 2022; Min et al., 2022). Therefore, the performance of LR-ASD using ResNet-18 to extract audio features is evaluated. While this encoder significantly increases the model parameters, it does not bring performance improvement, probably for similar reasons to the poor performance of ResNet in the visual feature encoders. Larger models may be prone to overfitting when extracting information from feature maps of small dimensions. Subsequently, in comparison with Light-ASD, LR-ASD's audio feature encoder is not only more lightweight but also demonstrates better performance. Finally, the performance of using 2D convolutions in audio blocks is validated. Due to the compact design of LR-ASD's encoder and the relatively lower computational complexity of 2D convolutions compared to 3D convolutions, the difference in the number of model parameters and FLOPs before and after splitting of 2D convolutions is relatively small. The experiments indicate that the performance of the audio encoder based on 2D convolutions is inferior to that of the audio encoder based on 1D convolutions achieved through the splitting of the former. Perhaps the audio feature maps lack a strong spatial logic similar to that of images, thus processing the MFCCs and temporal dimensions separately is more conducive to aggregating audio information.

4.5.5 Feature Fusion

Table 9 presents the performance of LR-ASD when fusing visual and audio features through different methods. In the task of active speaker detection, employing transformer-based multi-modal feature fusion is a common approach (Wang et al., 2024; Tao et al., 2021; Xiong et al., 2022). While LR-ASD achieves an mAP of nearly 94.5% using the transformer-based module (Tao et al., 2021) to fuse visual and audio features, it increases the number of parameters by 39%. Contrastingly, the element-wise addition employed in Light-ASD is the simplest and most lightweight feature fusion method, based on the assumption that both visual and audio features contribute equally to the prediction. However, previous research (Köpüklü et al., 2021) has demonstrated notable disparities in the performance of models relying solely on single-modal information for predictions in this task, suggesting that the contributions of visual and audio features are not equal. As a result, LR-ASD, using the element-wise addition, achieves an mAP of only 93.6%. Theoretically, the fusion method of concatenating features can maximize the preservation of information from different modalities, thus achieving better performance, as confirmed by subsequent experiments. To better utilize multi-modal features, LR-ASD, building upon feature concatenation, further refined audio-visual features using a simple channel attention module, resulting in an optimal mAP of 94.5%. Compared to the fusion method of element-wise addition, this method increases the number of parameters by 0.12 M in exchange for a 0.9% improvement in performance, which is acceptable.

4.5.6 Detector

The impact of detectors employing different methods to process audio-visual features on the performance of LR-ASD is presented in Table 10. When the detector directly inputs unprocessed audio-visual features into the FC layer for prediction, LR-ASD achieves an mAP of only 90.5%. When utilizing forward GRU for temporal modeling of audio-visual features, there is a 2.7% increase in mAP, demonstrating that the temporal context information of audio-visual features contributes to enhancing the performance of the active speaker detection model. Nevertheless, the forward GRU operates unidirectionally, resulting in an information imbal-

Table 9 Impact of feature fusion

Fusion	Params (M)	FLOPs (G)	mAP (%)
Transformer-based (Tao et al., 2021)	1.17	**0.51**	94.4
Add	**0.72**	**0.51**	93.6
Concate	0.77	**0.51**	94.0
LR-ASD	0.84	**0.51**	**94.5**

Bold represents the best result in the comparison methods

Table 10 Impact of detector

Detector	Params (M)	FLOPs (G)	mAP (%)
None	**0.70**	**0.51**	90.5
Forward GRU	0.76	**0.51**	93.2
Forward GRU (w channel attention)	0.76	**0.51**	93.4
Bidirectional GRU	0.82	**0.51**	94.0
Transformer (Vaswani et al., 2017)	1.49	**0.51**	93.0
LR-ASD	0.84	**0.51**	**94.5**

Bold represents the best result in the comparison methods

ance across the frames within the sequence. To this end, this study employs two temporal modeling approaches: bidirectional GRU and Transformer (Vaswani et al., 2017). These methods ensure that each frame in the sequence can incorporate information from the entire sequence to aid in prediction. Experiments demonstrate that the detector equipped with a bidirectional GRU outperforms that equipped with a Transformer. In this task, information from frames near the current detection frame is more beneficial for determining whether the candidate is speaking. Hence, the advantage of the Transformer, which ensures that all frames in the sequence have an equal opportunity to influence the current detection frame, is no longer significant. In contrast, the forgetting mechanism of the GRU enhances the informativeness of neighboring frames, rendering it a better choice for this task. Finally, building upon bidirectional GRU, LR-ASD adds a simple channel attention module to refine bidirectional temporal information, resulting in an optimal mAP of 94.5%. It is worth noting that the detector with the backward GRU is non-causal. To this end, we evaluate a causal detector composed of a forward GRU and a channel attention module. Compared to the Transformer-based causal detector, this detector not only improves model performance but also reduces the overall model parameters by nearly half.

4.5.7 Module Combination

We investigate the impact of each module in LR-ASD on the final performance in Table 11 and summarize the following findings. (a) Rows 1 and 2 in Table 11 confirm the previous conclusion (Köpüklü et al., 2021) that visual features have a more significant impact on the results than audio features in the task of active speaker detection. This phenomenon may be attributed to the circumstance that in multi-person scenarios, the visual information within the face sequence is capable of identifying the speaker, while the audio information can only determine whether someone is speaking, without specifically identifying the speaker. (b) Combining visual and audio features can significantly improve performance. Due to the varying contributions of visual and audio features, the proposed feature fusion module aims to fully exploit the potential of audio-visual features by dynamically assigning weights to visual and audio features, thereby further improving performance. (c) After incorporating the detector for temporal modeling of diverse features, the performance has been greatly improved, highlighting the importance of the detector module. (d) When all modules are applied together, the model achieves the highest mAP of 94.5%, indicating that each module in LR-ASD is indispensable.

4.5.8 Input Length

Table 12 illustrates the impact of training and testing LR-ASD with sequences of varying lengths on its performance. When the length of the input sequence is 1 frame, the mAP of LR-ASD is only 71.0%. The research indicates that an audio-visual episode lasting 5 s contains an average of 15 words (Cutts, 2020; Tauroza & Allison, 1990), whereas a segment of just 1 frame, which is approximately 0.04 s, may not even cover a complete word, resulting in a low mAP. As the number of input frames increases, the semantic information contained in the sequence becomes more abundant, consequently elevating LR-ASD's mAP to 93.2%. However, an increase in sequence length leads to a reduction in the number of video segments available for training. Therefore, when the input frame number increases from 100 to 150 frames, it

Table 11 Impact of module combination

#	Visual encoder	Audio encoder	Feature fusion	Detector	mAP (%)
1	✔				81.5
2		✔			51.0
3	✔	✔			89.7
4	✔	✔	✔		90.5
5	✔			✔	84.8
6		✔		✔	53.5
7	✔	✔		✔	93.6
8	✔	✔	✔	✔	**94.5**

Bold represents the best result in the comparison methods

Table 12 Impact of input length

Method	Video frames	mAP (%)
LR-ASD	1 (about 0.04 s)	71.0
	5 (about 0.2 s)	84.1
	10 (about 0.4 s)	87.6
	25 (about 1 s)	90.9
	50 (about 2 s)	92.2
	100 (about 4 s)	93.2
	150 (about 6 s)	92.0
	200 (about 8 s)	92.1
	Variable (about 1-10 s)	**94.5**

Table 13 Comparison of inference speed

Method	Video frames	Time (ms)	FPS
LoCoNet (Wang et al., 2024)	1 (about 0.04 s)	9.54	105
	1000 (about 40 s)	202.62	4935
	2000 (about 80 s)	412.81	4845
	4000 (about 160 s)	out of memory	
Light-ASD (Ours) (Liao et al., 2023)	1 (about 0.04 s)	4.49	223
	1000 (about 40 s)	96.04	10412
	2000 (about 80 s)	194.51	10282
	10000 (about 400 s)	1004.80	9952
	11000 (about 440 s)	out of memory	
LR-ASD	1 (about 0.04 s)	3.86	259
	1000 (about 40 s)	77.31	12935
	2000 (about 80 s)	154.93	12909
	10000 (about 400 s)	823.44	12144
	13000 (about 520 s)	1078.57	12053

does not yield a significant improvement in performance. In contrast, the experimental results confirm the conclusion of previous research (Tao et al., 2021) that using variable length sequences for training and testing can help the active speaker detection model achieve optimal performance. Finally, since LR-ASD is a non-causal model, the input data for real-time processing needs to contain lookahead information. Compared to the results (75.2%, 82.8%, and 87.9%) reported in TalkNet (Tao et al., 2021) for lookahead of 5, 10, and 25 frames, LR-ASD improves performance by 8.9%, 4.8%, and 3.0%, respectively. This indicates that the lightweight and low computational cost LR-ASD is a promising choice for real-time detection, particularly in resource-constrained environments.

4.5.9 Inference Speed

We evaluate Light-ASD (Liao et al., 2023), LR-ASD, and the state-of-the-art LoCoNet (Wang et al., 2024) on an NVIDIA RTX 3090 GPU to measure the model inference time and frames per second (FPS) under various numbers of input frames. The results are presented in Table 13. In the extreme case of single-frame input, LR-ASD achieves the fastest inference time, with only 3.86 ms, while the state-of-the-art LoCoNet reaches as high as 9.54 ms. As the number of input frames increases, the GPU becomes fully utilized. When the number of input frames reaches 2000, LR-ASD achieves an FPS of up to 12,909, which is approximately 2.7 times higher than that of LoCoNet. However, as the input frame number increases to 4000, the Video Random Access Memory (VRAM) of the 3090 GPU is insufficient to support inference for LoCoNet. In contrast, LR-ASD can utilize the same VRAM for inferring inputs of up to 13,000 frames. Furthermore, in comparison to Light-ASD presented in our conference publication, LR-ASD demonstrates not only a faster inference speed but also the capability to handle a larger number of input frames within the same VRAM. The factors contributing to this phenomenon may be attributed to two aspects. Firstly, LR-ASD possesses smaller parameters and computations. Secondly, LR-ASD's encoders, designed by a straight-barrel structure, require less VRAM during inference compared to Light-ASD's dual-path structure. Overall, in comparison to the state-of-the-art approach and lightweight method, LR-ASD not only sup-

ports longer input sequences under the same configuration but also has a higher inference speed.

4.6 Quantitative Analysis

Following the state-of-the-art methods (AVA (Roth et al., 2020), ASC (Alcázar et al., 2020), MAAS (Alcázar et al., 2021), TalkNet (Tao et al., 2021), ASDNet (Köpüklü et al., 2021), Light-ASD (Liao et al., 2023), LoCoNet (Wang et al., 2024)), we conduct a performance breakdown of LR-ASD on the benchmark AVA-ActiveSpeaker dataset based on the number and size of faces. Figure 9 presents the performance of various methods across diverse scenarios.

Figure 9a reports the impact of the number of visible faces in video frames on model performance. As the number of faces increases, the active speaker detection task becomes more challenging, and the performance of all methods degrades accordingly. Despite adopting a single-candidate input strategy to reduce computational complexity, LR-ASD demonstrates remarkable competitiveness across scenarios with varying numbers of visible faces. It outperforms the majority of methods employing multi-candidate input strategies (ASC, MAAS, and ASDNet), ranking second only to the state-of-the-art LoCoNet. This indicates that LR-ASD is capable of extracting refined and reliable audio-visual information from the current candidate, enabling accurate prediction without introducing relational context information between multiple candidates.

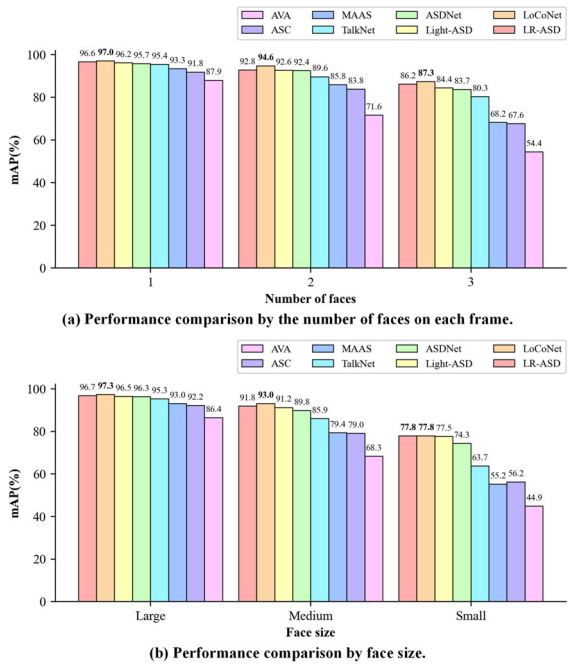


Fig. 9 Performance breakdown

Figure 9b illustrates the impact of varying face sizes on the performance of active speaker detection methods. Herein, the validation set is partitioned into three subsets based on the width of the detected faces: large (facial widths exceeding 128 pixels), medium (facial widths ranging between 64 and 128 pixels), and small (facial widths below 64 pixels). Although the performance of all methods decreases with decreasing face size, the performance of LR-ASD remains second only to the state-of-the-art LoCoNet. Furthermore, LR-ASD achieves a state-of-the-art mAP of 77.8% in scenarios featuring small-sized faces, aligning with the performance of LoCoNet.

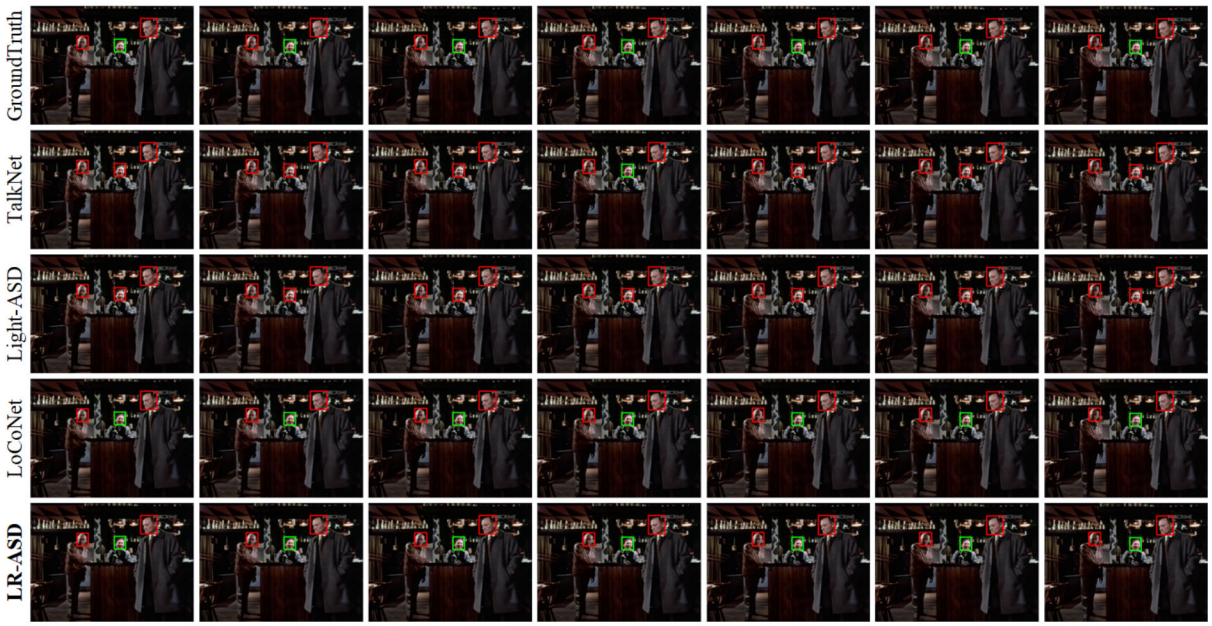
In summary, LR-ASD, as a lightweight solution for active speaker detection, achieves performance close to the state-of-the-art LoCoNet in six scenarios using only a minimal number of parameters, demonstrating its robustness.

4.7 Qualitative Analysis

Figure 10 illustrates the prediction results of four active speaker detection methods, TalkNet, Light-ASD, LoCoNet, and LR-ASD, on the AVA-ActiveSpeaker dataset along with the corresponding groundtruth labels. Figure 10a, b show challenging scenes featuring 2 and 3 small-sized faces, respectively. In Fig. 10a, LR-ASD accurately predicts the active speaker in the full-shot (Rao et al., 2020) scene, in contrast to the false positive results obtained by TalkNet and the false negative outcomes of Light-ASD and LoCoNet. Additionally, both LR-ASD and the state-of-the-art LoCoNet accurately identify the active speaker in all frames of Fig. 10b, whereas the second-ranked TalkNet correctly identifies the active speaker in only one frame. Overall, LR-ASD enhances the competitiveness of active speaker detection methods employing a single-candidate input strategy in challenging scenarios with multiple small-sized faces. This can be attributed to the compact and ingenious design of LR-ASD, which facilitates the model in learning common features of speaking behavior.

Subsequently, we summarize 5 categories of failure predictions in LR-ASD and present them in Fig. 11. (a) Error annotation. The dataset (Roth et al., 2020) contains a few erroneous annotations, particularly in dubbed movies. (b) Lip movement. Someone is speaking off-screen, while the on-screen characters exhibit lip movements such as eating or yawning. (c) Blurred face. Some videos have lower quality, with the entire frame appearing blurry. (d) Occluded face. The region around the speaker's lip is occluded, resulting in the loss of relevant visual information. (e) Head pose. The speaker faces the camera sideways, making it difficult to observe changes in the lip region. The categories (a) and (c) of failure predictions can be solved by improving data quality. As for (b), (d), and (e), it is difficult for the model to make accurate predictions solely based on limited facial informa-

(a) Sample video 1 features a two-person scene with small-sized faces.



(b) Sample video 2 features a three-person scene with small-sized faces.

Fig. 10 Comparison of results for TalkNet (Tao et al., 2021), Light-ASD (Liao et al., 2023), LoCoNet (Wang et al., 2024), and LR-ASD in challenging scenarios within the validation set of the AVA-ActiveSpeaker dataset. Red boxes indicate inactive speakers and green boxes indicate active speakers (Color figure online)

tion. It may need to be solved by adding information on the speaker's body or environment, but this will undoubtedly significantly increase the computational burden. Therefore, how to maintain the model lightweight while further improving performance in active speaker detection tasks is a problem worth further exploration.

4.8 Discussion

4.8.1 Scene Scalability

To validate the feasibility of the proposed methods in more challenging scenarios, we evaluate the performance of LR-ASD and Light-ASD using the EasyCom dataset (Donley

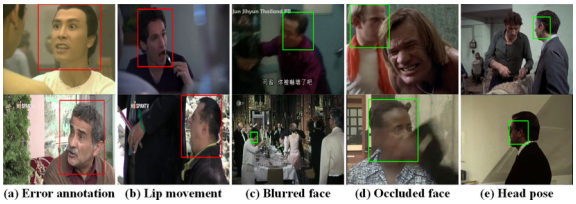


Fig. 11 Categories of failure predictions. The red and green boxes respectively indicate inactive and active speakers in the labels (Color figure online)

Table 14 Comparison of mAP (%) on the EasyCom dataset (Donley et al., 2021)

Method	Training set		
	AVA	EasyCom	AVA & EasyCom
Light-ASD (Ours)	53.1	84.9	90.1
LR-ASD	**57.8**	**86.4**	**90.9**

Bold represents the best result in the comparison methods

et al., 2021). This is a multi-modal augmented reality (AR) dataset comprising approximately 6 h of egocentrically recorded conversational video using AR glasses in a noisy restaurant-like room to address the cocktail party problem. The experimental results presented in Table 14 indicate that LR-ASD consistently outperforms Light-ASD across various experimental settings. Overall, LR-ASD trained solely on the AVA-ActiveSpeaker dataset not only comprehensively outperforms Light-ASD on conventional active speaker detection datasets (AVA-ActiveSpeaker, Talkies, Columbia, and RealVAD), but also maintains a performance advantage of 4.7% on the challenging augmented reality dataset. This indicates that the improvements made by LR-ASD to Light-ASD in feature fusion and temporal modeling facilitate the extraction of more refined speaker features by LR-ASD, thereby demonstrating more robust performance across different datasets. Concurrently, LR-ASD pre-trained on the AVA-ActiveSpeaker dataset exhibits the capability for rapid adaptation to novel scenes through fine-tuning.

4.8.2 Task scalability

We utilize the largest in-the-wild facial expression recognition dataset, FERV39k (Wang et al., 2022), to assess the feasibility of extending the proposed active speaker detection methods to other tasks. This dataset comprises nearly 39k video clips labeled with 7 basic facial expressions. Since the FERV39K dataset does not contain audio data and the facial expression recognition task requires a unique prediction of the input face sequence, Light-ASD and LR-ASD remove the audio feature encoder and add global average pooling processing sequence temporal information after the detector to meet the task requirements. The experimental results in Table 15 indicate that LR-ASD performs significantly better than classical baseline methods in the facial expression recognition task, including C3D utilizing 3D convolutions and R(2+1)D constructed with combinations of 2D and 1D convolutions. As for the state-of-the-art method (Li et al., 2023) designed for this task, LR-ASD achieves comparable performance with only about 3% of its parameters, with unweighted average recall (UAR) equal and weighted average recall (WAR) lagging by only 2.8%. It is worth noting that fine-tuning models pre-trained on the AVA-ActiveSpeaker dataset does not lead to performance improvements in this new task, possibly attributable to significant differences in the feature focus between the two tasks. Among these tasks, the facial expression recognition task emphasizes changes in facial features across the entire sequence, while the active speaker detection task prioritizes the match of lip movements with the corresponding audio. Therefore, the model initialized with weights pre-trained on the AVA-ActiveSpeaker dataset still needs to learn essential features from scratch for facial expression recognition. Finally, compared to Light-ASD, LR-ASD exhibits better scalability in the facial expression recognition task, highlighting its effectiveness in extending to tasks involving facial analysis.

Table 15 Comparison of performance on the FERV39k dataset (Wang et al. 2022)

Method	Avenue	Params (M)	UAR (%)	WAR (%)
C3D (Tran et al., 2015)	ICCV'15	78.02	22.7	31.7
R(2+1)D (Tran et al., 2018)	CVPR'18	33.18	31.6	41.3
M3DFEL (Wang et al., 2023)	CVPR'23	–	**35.9**	47.7
IAL (Li et al., 2023)	AAAI'23	19.08	35.8	**48.5**
Light-ASD (Ours)	CVPR'23	0.74	33.8	44.4
Light-ASD* (Ours)	CVPR'23	0.74	33.9	44.9
LR-ASD	–	**0.51**	34.8	45.8
LR-ASD*	–	**0.51**	35.8	45.7

Bold represents the best result in the comparison methods
*:pre-trained on the AVA-ActiveSpeaker dataset

5 Conclusion

In this study, a lightweight and robust network for active speaker detection, named LR-ASD, is proposed. The key lightweight features of LR-ASD include inputting a single candidate, splitting 2D and 3D convolutions to separately extract audio and visual features, and employing simple modules for multi-modal feature fusion and temporal modeling. The results on the benchmark dataset (Roth et al., 2020) indicate that LR-ASD reduces the model parameters by 97.6% and FLOPs by 89.5% compared with the state-of-the-art method (Wang et al., 2024), with mAP lagging by only 0.7%. Subsequently, in cross-dataset testing without fine-tuning on three public datasets (Alcázar et al., 2021; Chakravarty & Tuytelaars, 2016; Beyan et al., 2020), LR-ASD achieves state-of-the-art performance, demonstrating its outstanding robustness. Finally, LR-ASD has exhibited good scalability in augmented reality environments (Donley et al., 2021) and the facial expression recognition task (Wang et al., 2022).

Acknowledgements This work is supported in part by the National Natural Science Foundation of China (62302324, 62072319 and 62272329); in part by the National Key Research and Development Program of China (2023YFB3308300); in part by the Sichuan Science and Technology Program (2023YFQ0022).

Data Availibility All experiments are performed on publicly available dataset: AVA-ActiveSpeaker (https://research.google.com/ava/download.html\#ava_active_speaker_download), Talkies (https://filedn.com/l0kNCNuXuEq70c3iUHsXxJ7/Talkies/), Columbia (https://www.youtube.com/watch?v=6GzxbrO0DHM\&t=2s), RealVAD (https://zenodo.org/record/3928151), TalkSet (https://github.com/TaoRuijie/TalkNet-ASD), EasyCom (https://github.com/facebookresearch/EasyComDataset/releases), and FERV39k (https://github.com/wangyanckxx/FERV39k?tab=readme-ov-file).

References

Alcázar, J. L., Caba, F., Mai, L., Perazzi, F., Lee, J. -Y., Arbeláez, P., & Ghanem, B. (2020). Active speakers in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12465–12474).

Alcázar, J. L., Caba, F., Thabet, A. K., & Ghanem, B. (2021). Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 265–274).

Afouras, T., Chung, J. S., & Zisserman, A. (2018). Lrs3-ted: A large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.

Alcázar, J. L., Cordes, M., Zhao, C., & Ghanem, B. (2022). End-to-end active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Part XXXVII* (pp. 126–143). Springer.

Afouras, T., Owens, A., Chung, J. S., & Zisserman, A. (2020). Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European conference, Part XVIII 16* (pp. 208–224). Springer.

Arandjelovic, R., & Zisserman, A. (2018). Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 435–451).

Ban, Y., Alameda-Pineda, X., Girin, L., & Horaud, R. (2021). Variational Bayesian inference for audio-visual tracking of multiple speakers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(5), 1761–1776.

Beyan, C., Shahid, M., & Murino, V. (2020). Realvad: A real-world dataset and a method for voice activity detection by body motion analysis. *IEEE Transactions on Multimedia, 23*, 2071–2085.

Cutler, R., & Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In *2000 IEEE international conference on multimedia and expo* (vol. 3, pp. 1589–1592). IEEE.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Chung, J. S. (2019). Naver at activitynet challenge 2019–task b active speaker detection (ava). arXiv preprint arXiv:1906.10555.

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.

Chakravarty, P., & Tuytelaars, T.: Cross-modal supervision for learning active speaker detection in video. In *Computer Vision–ECCV 2016: 14th European conference, Part V 14* (pp. 285–301). Springer.

Cutts, M. (2020). *Oxford guide to plain English*. Oxford University Press.

Chung, J. S., & Zisserman, A. (2017). Out of time: Automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13* (pp. 251–263). Springer.

Chakravarty, P., Zegers, J., Tuytelaars, T., & Van hamme, H. (2016). Active speaker detection with audio-visual co-training. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 312–316).

Datta, G., Etchart, T., Yadav, V., Hedau, V., Natarajan, P., & Chang, S. -F. (2022). 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)* (pp. 4568–4572). IEEE.

Duan, H., Liao, J., Lin, L., El Saddik, A., & Cai, W. (2024). Meetor: A human-centered automatic video editing system for meeting recordings. *ACM Transactions on Multimedia Computing, Communications and Applications, 20*(9), 1–23.

Duan, H., Liao, J., Lin, L., & Cai, W. (2022). 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* (pp. 43–49).

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366.

Donley, J., Tourbabin, V., Lee, J. -S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V. K., & Mehra, R. (2021). Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. arXiv preprint arXiv:2107.04174.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13733–13742).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778)

Jati, A., & Georgiou, P. (2019). 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.

Jiang, Y., Tao, R., Pan, Z., & Li, H. (2023). Target active speaker detection with audio-visual cues. arXiv preprint arXiv:2305.12831.

Krawczyk, M., & Gerkmann, T. (2014). Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(12), 1931–1940.

Köpüklü, O., Taseska, M., & Rigoll, G. (2021). 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* (pp. 1193–1203).

Liu, Z.-S., Courant, R., & Kalogeiton, V. (2024). Funnynet-W: Multimodal learning of funny moments in videos in the wild. *International Journal of Computer Vision, 132*, 2885–2906.

Liao, J., Duan, H., Feng, K., Zhao, W., Yang, Y., & Chen, L. (2023). A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 22932–22941).

Liao, J., Duan, H., Li, X., Xu, H., Yang, Y., Cai, W., Chen, Y., & Chen, L. (2020) Occlusion detection for automatic video editing. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2255–2263) (2020)

Liao, J., Duan, H., Zhao, W., Yang, Y., & Chen, L. (2022). A light weight model for video shot occlusion detection. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3154–3158). IEEE.

Liao, J., Duan, H., Zhao, W., Feng, K., Yang, Y., & Chen, L. (2024). A video shot occlusion detection algorithm based on the abnormal fluctuation of depth information. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(3), 1627–1640.

Li, H., Niu, H., Zhu, Z., & Zhao, F. (2023). Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, (vol. 37, pp. 67–75).

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(2), 198–213.

Moattar, M. H., & Homayounpour, M. M.: A simple but efficient real-time voice activity detection algorithm. In *2009 17th European signal processing conference* (pp. 2549–2553). IEEE.

Min, K., Roy, S., Tripathi, S., Guha, T., & Majumdar, S. (2022). Learning long-term spatial-temporal graphs for active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Part XXXV* (pp. 371–387). Springer.

Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D., & Jensen, J. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 1368–1396.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689–696).

Owens, A., & Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*. (pp. 631–648).

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing, 13*(2), 206–219.

Planamente, M., Plizzari, C., Peirone, S. A., Caputo, B., & Bottino, A. (2024). Relative norm alignment for tackling domain shift in deep multi-modal classification. *International Journal of Computer Vision, 132*, 2618–2638.

Qian, X., Brutti, A., Lanz, O., Omologo, M., & Cavallaro, A. (2021). Audio-visual tracking of concurrent speakers. *IEEE Transactions on Multimedia, 24*, 942–954.

Qiao, M., Liu, Y., Xu, M., Deng, X., Li, B., Hu, W., & Borji, A. (2024). Joint learning of audio-visual saliency prediction and sound source localization on multi-face videos. *International Journal of Computer Vision, 132*(6), 2003–2025.

Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 1021–1028). IEEE.

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., & Xi, Z. (2020). 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)* (pp. 4492–4496). IEEE.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10428–10436).

Ravanelli, M., Parcollet, T., & Bengio, Y. (2019). The Pytorch-Kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6465–6469). IEEE

Ramırez, J., Segura, J. C., Benıtez, C., De La Torre, A., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication, 42*(3–4), 271–287.

Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., & Lin, D. (2020). A unified framework for shot type classification based on subject centric lens. In *Computer Vision–ECCV 2020: 16th European Conference, Part XI 16* (pp. 17–34). Springer.

Shahid, M., Beyan, C., & Murino, V. (2019). Comparisons of visual activity primitives for voice activity detection. In *Image analysis and processing–ICIAP 2019: 20th international conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20* (pp. 48–59). Springer.

Shahid, M., Beyan, C., & Murino, V. (2019). Voice activity detection by upper body motion analysis and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 0–0).

Shahid, M., Beyan, C., & Murino, V. (2021). S-vvad: Visual voice activity detection by motion segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision* pp. 2332–2341 (2021)

Slaney, M., & Covell, M. (2000). Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. *Advances in Neural Information Processing Systems* **13**.

Son Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6447–6456).

Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., & Darrell, T. (2005). Visual speech recognition with loosely synchronized feature streams. In *Tenth IEEE international conference on computer vision (ICCV'05) volume 1* (vol. 2, pp. 1424–1431). IEEE.

Sharma, R., & Narayanan, S. (2022). Unsupervised active speaker detection in media content using cross-modal information. arXiv preprint arXiv:2209.11896.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics, 11*(1), 90–105.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Truong, T.-D., Duong, C.N., Pham, H.A., Raj, B., Le, N., & Luu, K. (2021). The right to talk: An audio-visual transformer approach. In:

Proceedings of the IEEE/CVF international conference on computer vision (pp. 1105–1114).

Tesema, F.B., Lin, Z., Zhu, S., Song, W., Gu, J., Wu, H.: End-to-end audiovisual feature fusion for active speaker detection. In: Fourteenth International Conference on Digital Image Processing (ICDIP 2022), vol. 12342, pp. 681–688 (2022). SPIE

Tao, R., Pan, Z., Das, R. K., Qian, X., Shou, M. Z., & Li, H. (2021). Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3927–3935).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459).

Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O., & Ranjan, A. (2023). Mobileone: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7907–7917).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* **30**

Wang, X., Cheng, F., & Bertasius, G. (2024). Loconet: Long-short context network for active speaker detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18462–18472).

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018). Speaker diarization with lstm. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5239–5243). IEEE.

Wang, H., Li, B., Wu, S., Shen, S., Liu, F., Ding, S., & Zhou, A. (2023). Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17958–17968).

Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., & Zhang, W. (2022). Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20922–20931).

Wuerkaixi, A., Zhang, Y., Duan, Z., & Zhang, C. (2022). Rethinking audio-visual synchronization for active speaker detection. In *2022 IEEE 32nd international workshop on machine learning for signal processing (MLSP)* (pp. 01–06). IEEE.

Xiong, J., Zhou, Y., Zhang, P., Xie, L., Huang, W., & Zha, Y. (2022) Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Transactions on Multimedia* (pp. 1–14).

Zhang, Y., Liang, S., Yang, S., Liu, X., Wu, Z., & Shan, S. (2021a). Ictcas-ucas-tal submission to the ava-activespeaker task at activitynet challenge 2021. *The ActivityNet Large-Scale Activity Recognition Challenge, 1*(3), 4.

Zhang, Y., Liang, S., Yang, S., Liu, X., Wu, Z., Shan, S., & Chen, X. (2021b). Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3964–3972).

Zhang, Y. -H., Xiao, J., Yang, S., & Shan, S. (2019). Multi-task learning for audio-visual active speaker detection. The ActivityNet Large-Scale Activity Recognition Challenge (pp. 1–4).

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848–6856).