四川大學 SICHUAN UNIVERSITY
香港中文大學（深圳） The Chinese University of Hong Kong, Shenzhen
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

A Light Weight Model for Active Speaker Detection

Junhua Liao[1], Haihan Duan[2], Kanghui Feng[1], Wanbing Zhao[1], Yanbing Yang[1,3], Liangyin Chen[1,3]
1. College of Computer Science, Sichuan University, Chengdu, China
2. The Chinese University of Hong Kong, Shenzhen, China
3. The Institute for Industrial Internet Research, Sichuan University, Chengdu, China

Motivation

- Active speaker detection is a challenging task, with the aim to detect who is speaking in one or more speaker scenarios. This task has received considerable attention because it is crucial in many applications.
- Existing studies have attempted to improve the performance by inputting multiple candidate information and designing complex models. Although these methods have achieved excellent performance, their high memory and computational power consumption render their application to resource-limited scenarios difficult.

Contribution

- This study proposed a lightweight active speaker detection architecture and a novel loss function designed for training.
- Experimental results on the AVA-ActiveSpeaker dataset reveal that the proposed framework achieves competitive mAP performance (94.1% vs. 94.2%), while the resource costs are significantly lower than the state-of-the-art method, particularly in model parameters (1.0M vs. 22.5M, approximately 23x) and FLOPs (0.6G vs. 2.6G, approximately 4x).
- Ablation studies, cross-dataset testing, and qualitative analysis demonstrate the good robustness of the proposed method.

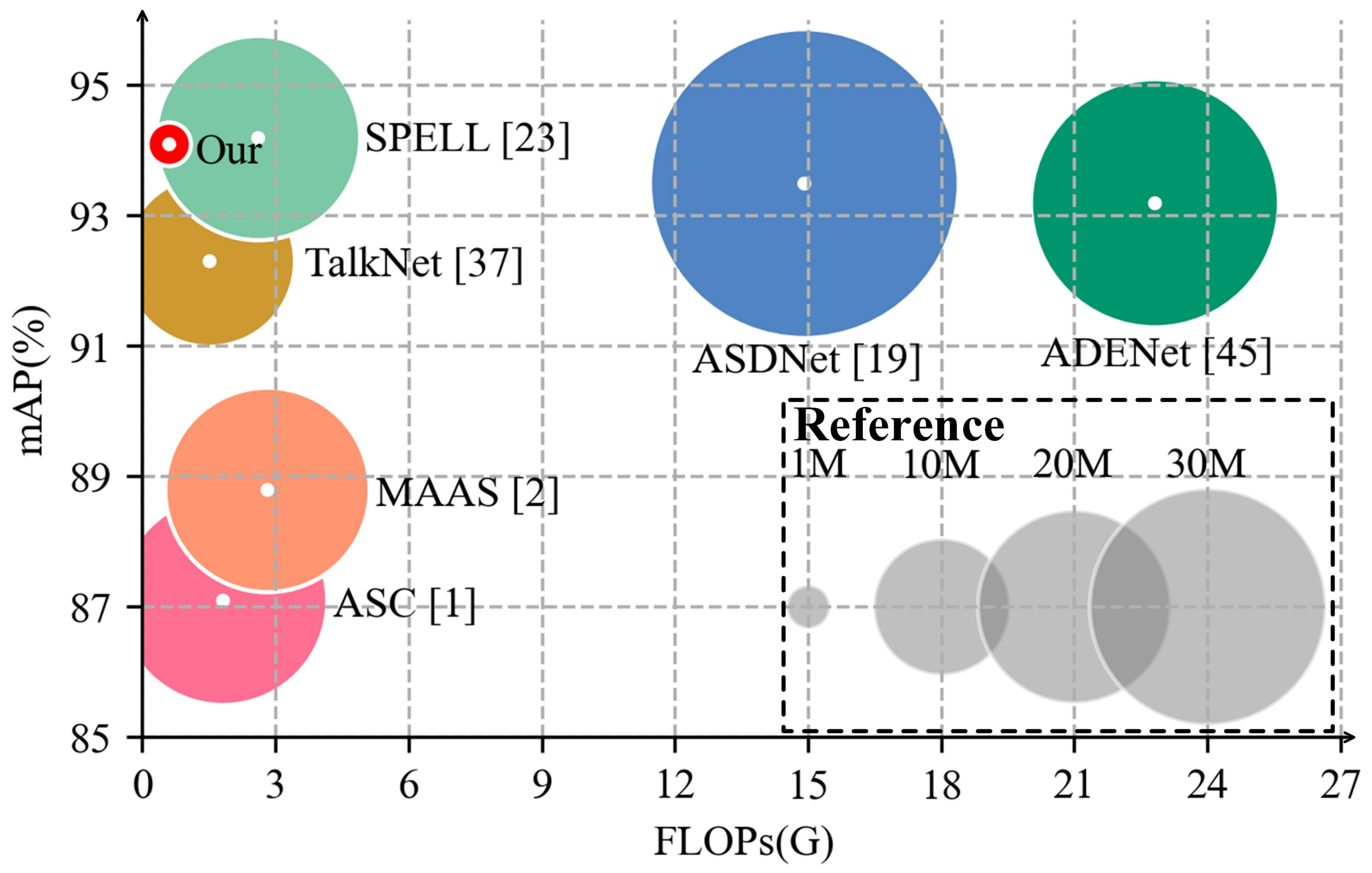

Figure 1: mAP vs. FLOPs, size ∝ parameters.

Proposed Method

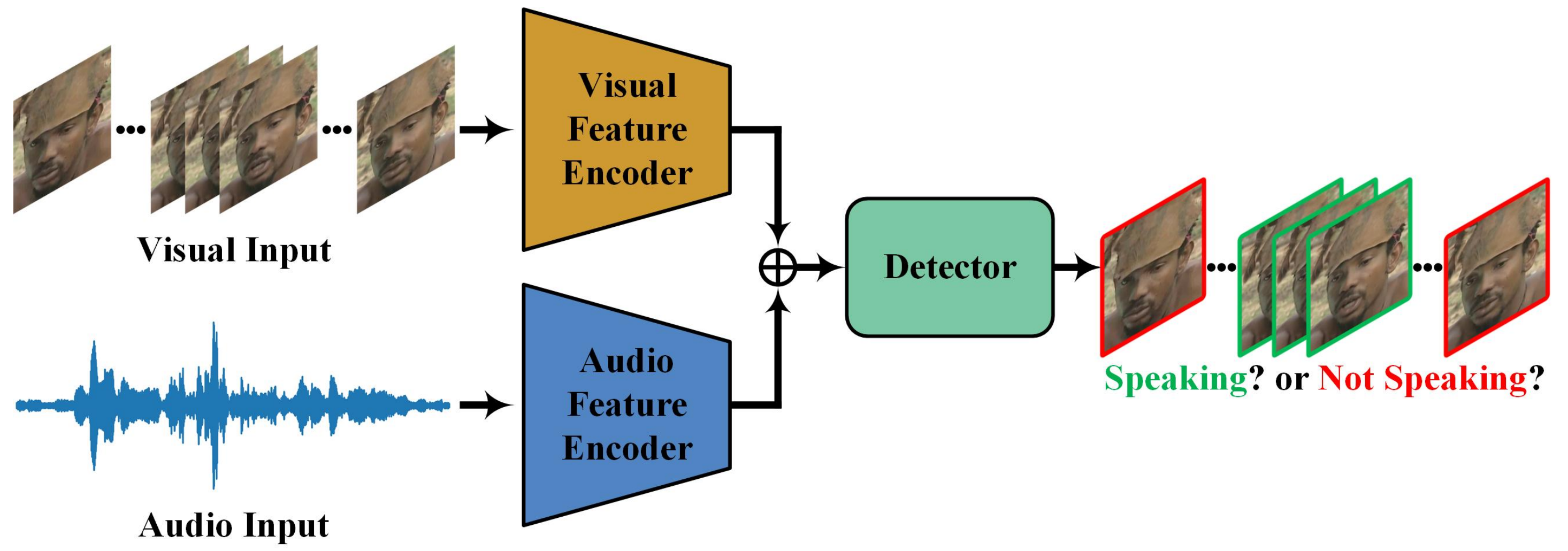

Figure 2: Overview of the proposed lightweight framework.

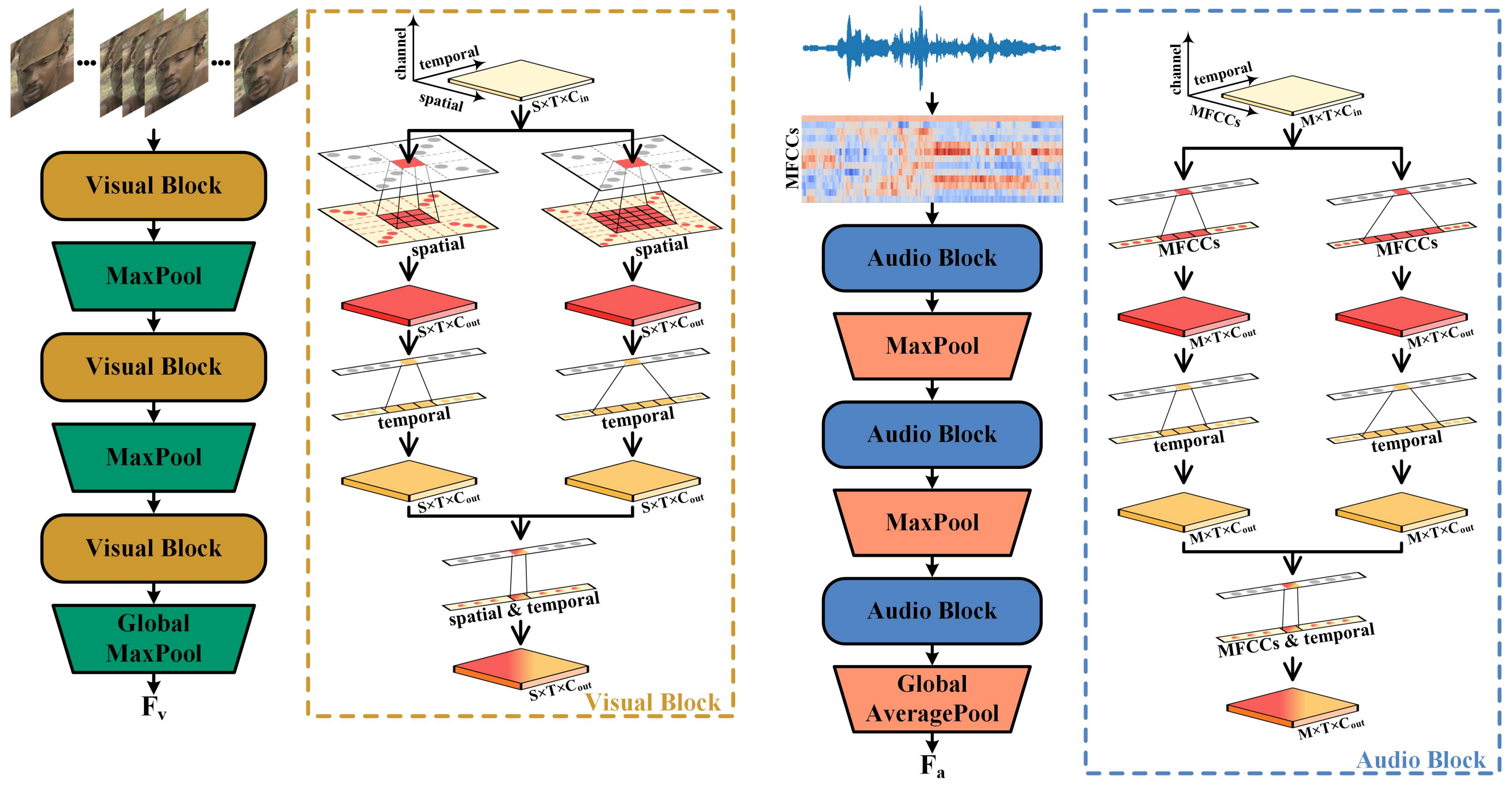

Figure 3: The architecture of the visual feature encoder and audio feature encoder.

This architecture has been improved in three aspects:

- **Single input**: inputting a single candidate face sequence with the corresponding audio;
- **Feature extraction**: splitting the 3D convolution of visual feature extraction into 2D and 1D convolutions to extract spatial and temporal information, respectively, and splitting the 2D convolution for audio feature extraction into two 1D convolutions to extract the frequency and temporal information;
- **Cross-modal modeling**: using GRU with less calculation, instead of complex attention modules, for cross-modal modeling.

Experiment

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

Table 1: Performance comparison on the validation set of AVA-ActiveSpeaker dataset.

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

Table 2: Performance comparison on the Columbia dataset.

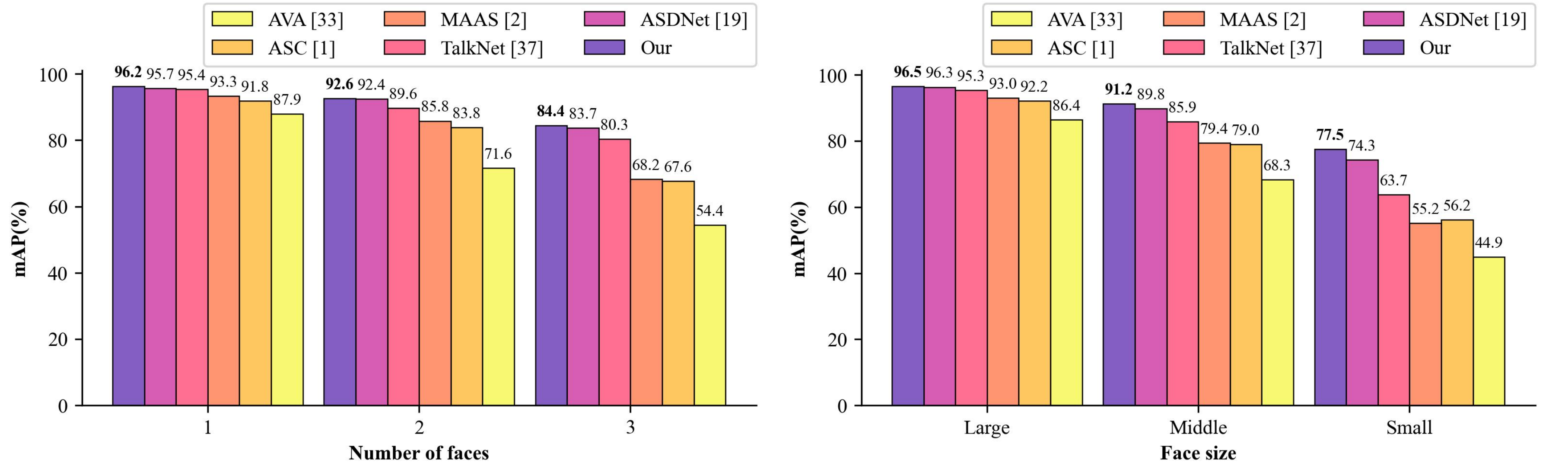

Figure 4: Qualitative Analysis.

Conclusion

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

Project page: https://github.com/Junhua-Liao/Light-ASD