

# A Light Weight Model for Active Speaker Detection

**Junhua Liao<sup>1</sup>, Haihan Duan<sup>2,3</sup>, Kanghui Feng<sup>1</sup>, Wanbing Zhao<sup>1</sup>, Yanbing Yang<sup>1,3</sup>, Liangyin Chen<sup>1,3</sup>**

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

**CVPR 2023 THU-PM-222**

# Highlights

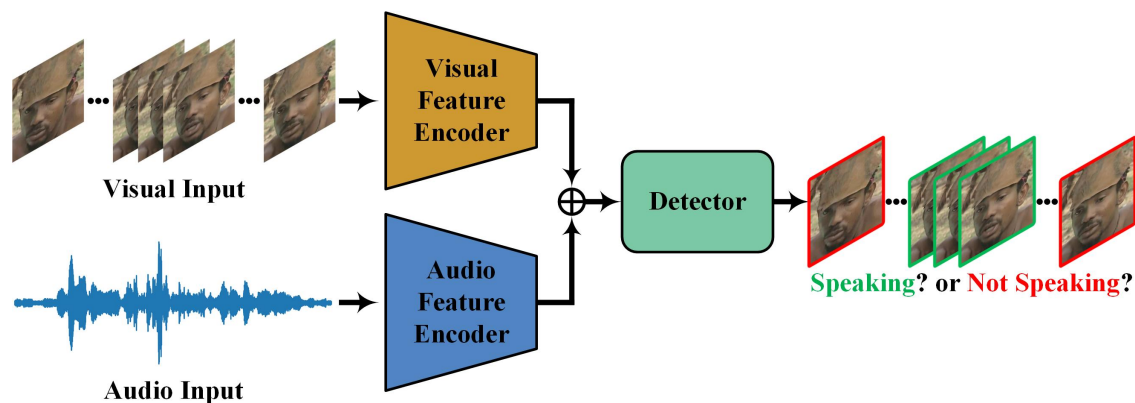


Figure 1. Overview of the proposed framework.

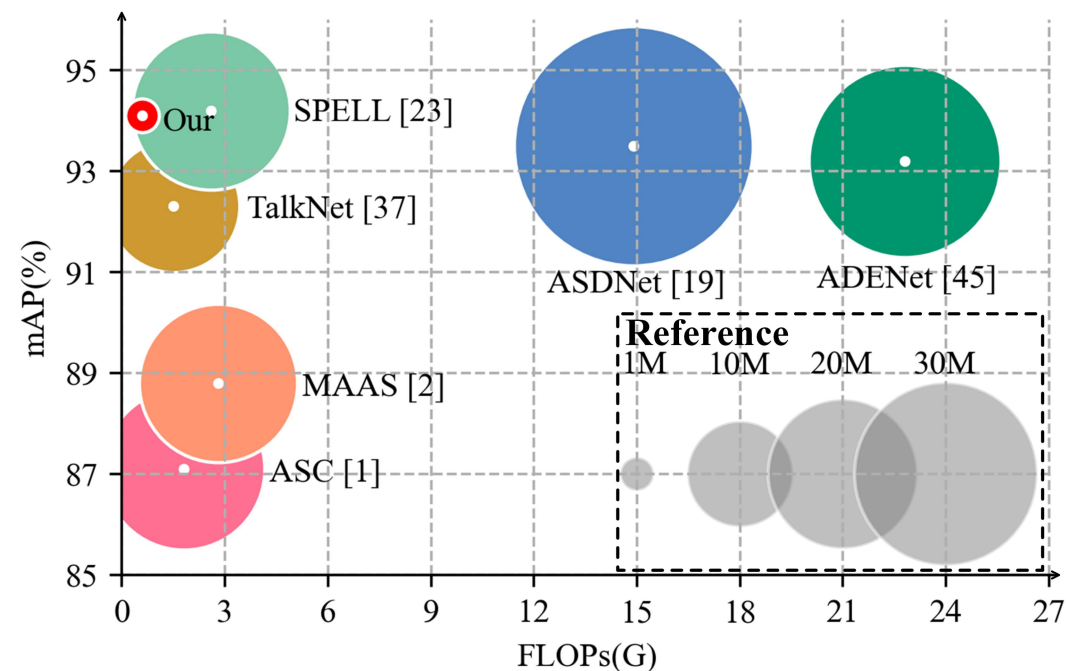


Figure 2. mAP vs. FLOPs, size  $\propto$  parameters.

# Contributions

- A lightweight design is developed from the three aspects of information input, feature extraction, and cross-modal modeling; subsequently, a lightweight and effective end-to-end active speaker detection framework is proposed. In addition, a novel loss function is designed for training.
- Experiments on *AVA-ActiveSpeaker*, a benchmark dataset for active speaker detection released by Google, reveal that the proposed method is comparable to the state-of-the-art method, while still reducing model parameters by 95.6% and FLOPs by 76.9%.
- Ablation studies, cross-dataset testing, and qualitative analysis demonstrate the state-of-the-art performance and good robustness of the proposed method.

# Encoders

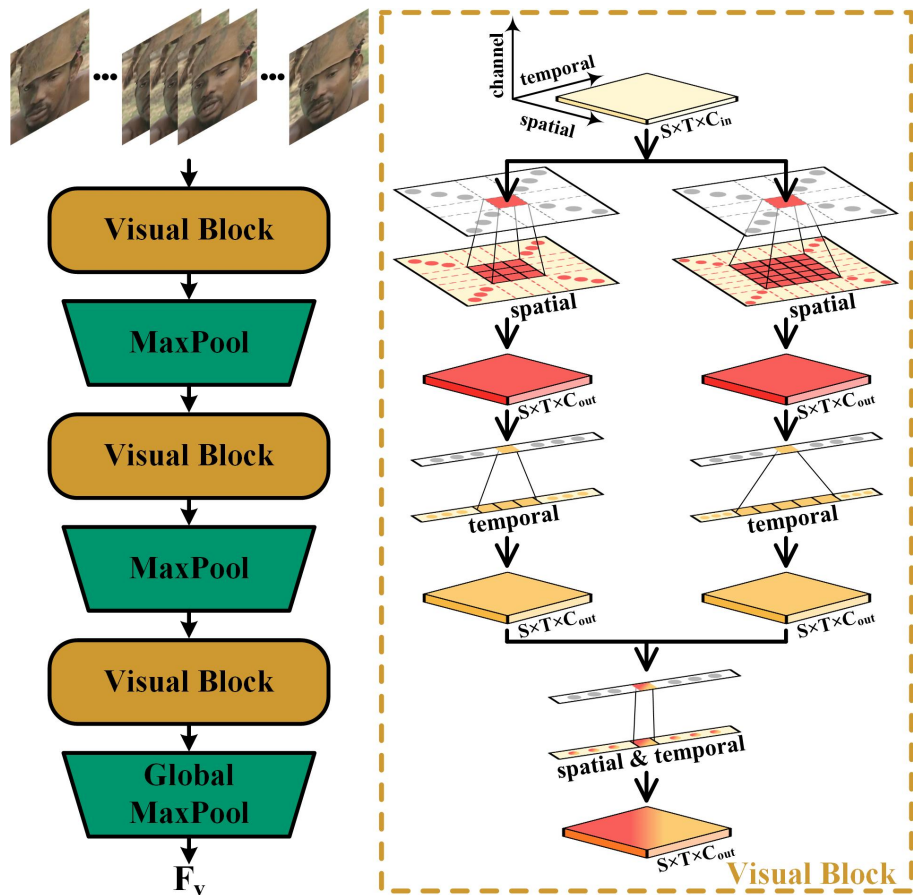


Figure 3. The architecture of visual feature encoder.

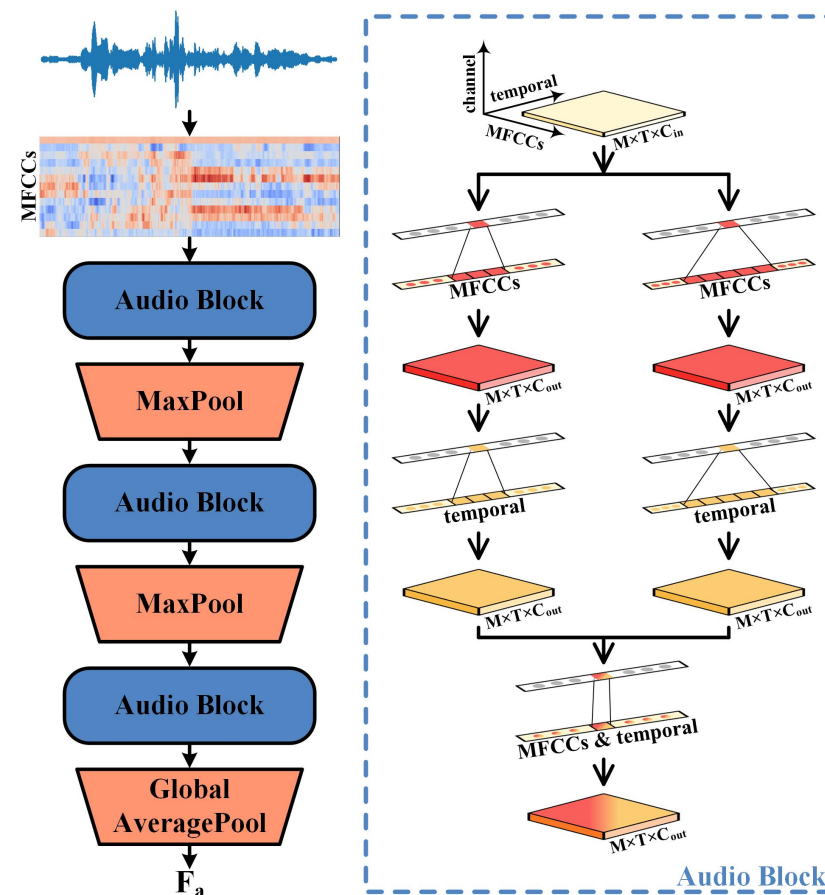


Figure 4. The architecture of the audio feature encoder.



# Detector & Loss

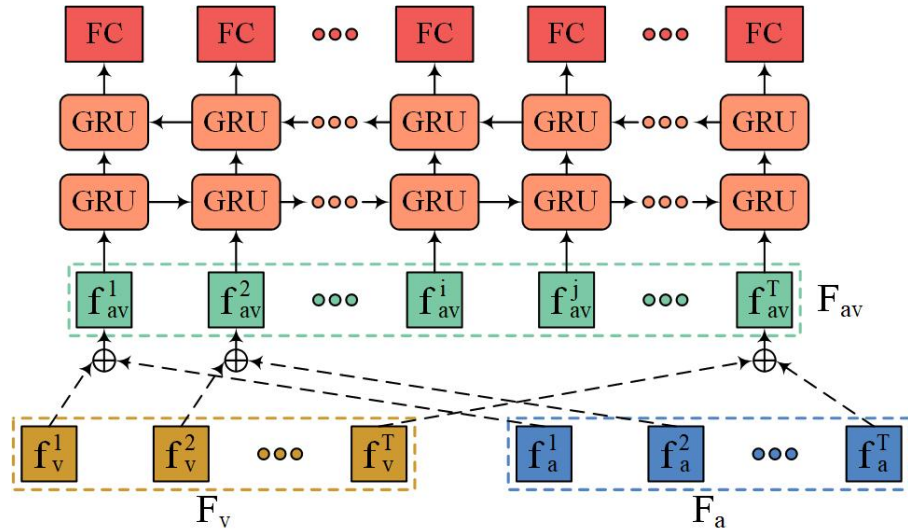


Figure 5. The architecture of the detector.

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

$$R = R_0 - \alpha E \cdots \cdots \cdots (2)$$

$$l = -\frac{1}{T} \sum_{i=1}^T (g^i \log(p_s^i) + (1 - g^i) \log(1 - p_s^i)) \cdots (3)$$

$$L_{asd} = l_{av} + \lambda l_v \cdots \cdots \cdots (4)$$

# Experiments

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	<b>94.2</b>
<b>Our Method</b>	✓	✗	✓	<b>1.0</b>	<b>0.6(0.2×3)</b>	94.1

*Table 1. Performance comparison for methods on the validation set of the AVA-ActiveSpeaker dataset.*

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

*Table 2. Comparison of F1-Score (%) on the Columbia dataset.*

# Ablation Studies

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

Table 3. Impact of convolutional kernel size.

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

Table 4. Impact of visual feature encoder.

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

Table 5. Impact of audio feature encoder.

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

Table 6. Impact of the detector.

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

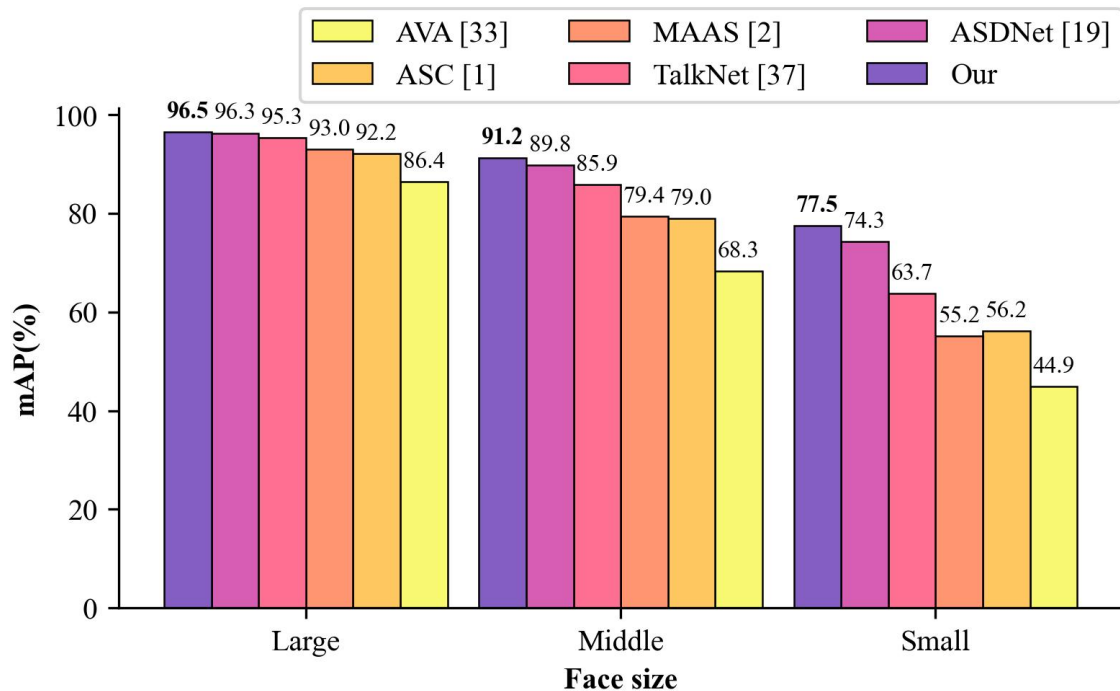
Table 7. Impact of the loss function.

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

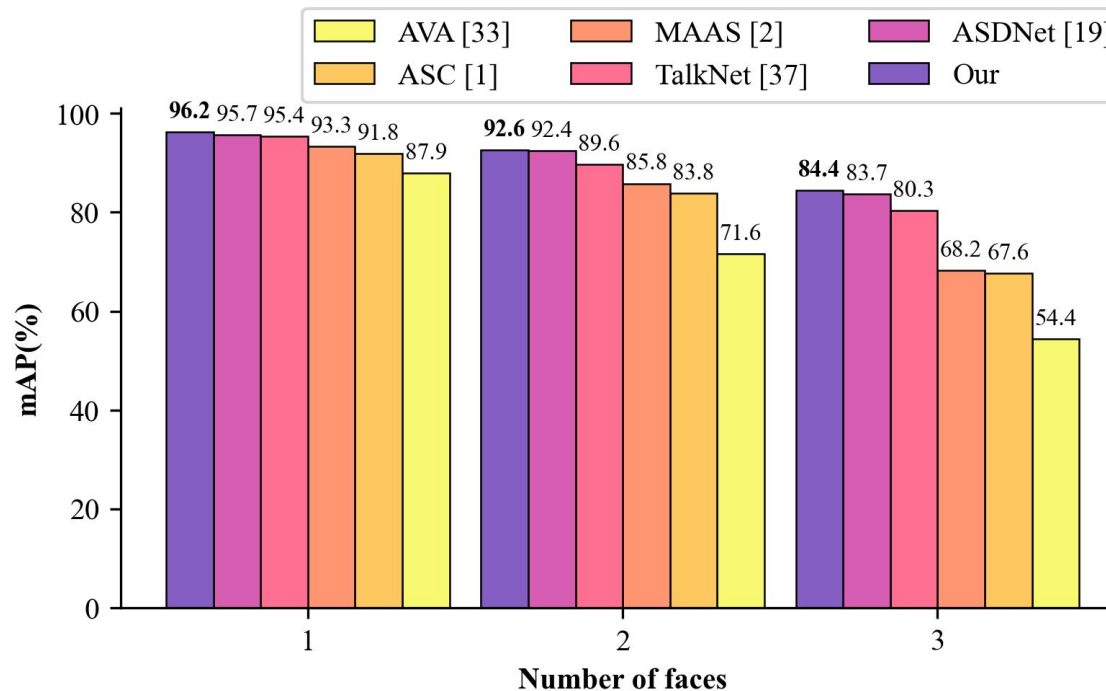
Table 8. Impact of the number of frames on the detection speed.



# Qualitative Analysis



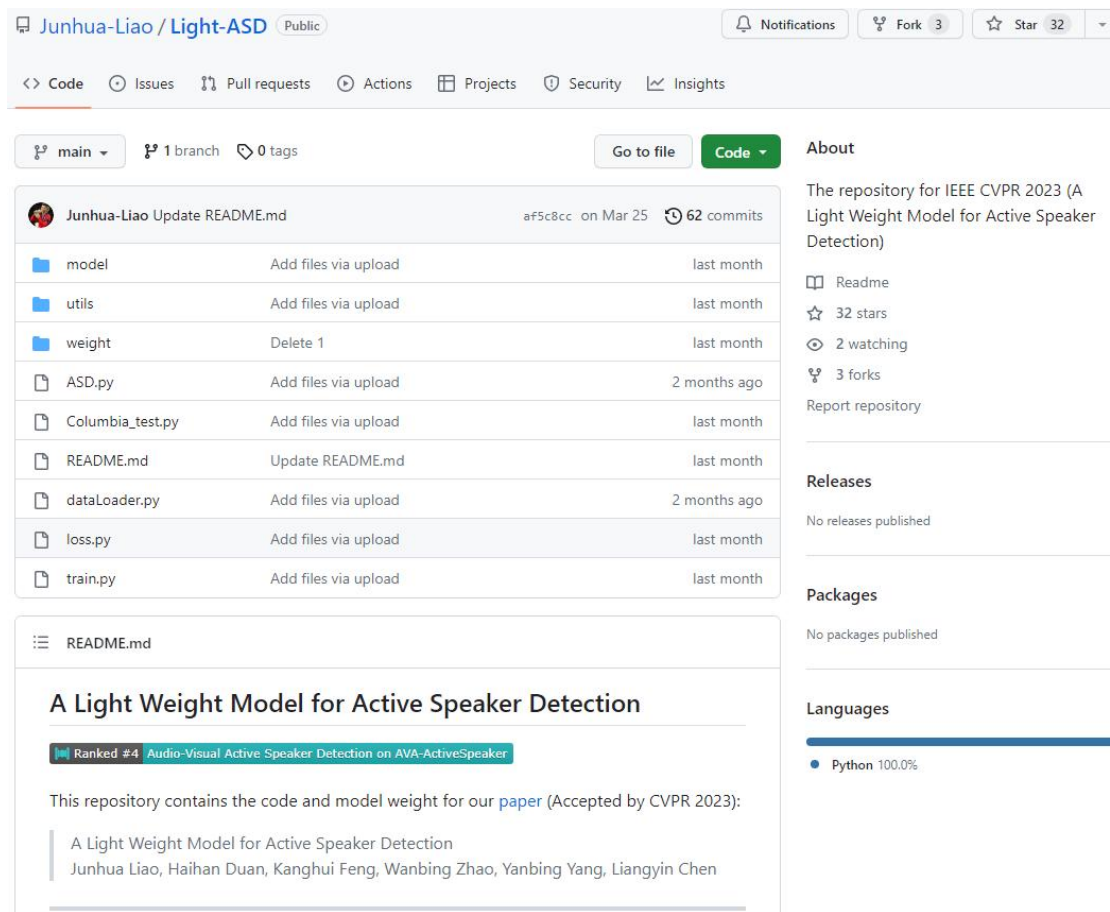
*Performance comparison by face size.*



*Performance comparison by the number of faces on each frame.*



# Project Page



The screenshot shows the GitHub repository page for Junhua-Liao / Light-ASD. The repository is public and has 32 stars, 3 forks, and 2 watchers. It contains a README.md file, a model directory, a utils directory, a weight directory, and several Python files (ASD.py, Columbia\_test.py, dataLoader.py, loss.py, train.py). The README.md file is expanded, showing the title "A Light Weight Model for Active Speaker Detection" and a description: "This repository contains the code and model weight for our paper (Accepted by CVPR 2023): A Light Weight Model for Active Speaker Detection by Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen".

Junhua-Liao / Light-ASD Public

Notifications Fork 3 Star 32

Code Issues Pull requests Actions Projects Security Insights

main 1 branch 0 tags Go to file Code About

Junhua-Liao Update README.md af5c8cc on Mar 25 62 commits

model	Add files via upload	last month
utils	Add files via upload	last month
weight	Delete 1	last month
ASD.py	Add files via upload	2 months ago
Columbia_test.py	Add files via upload	last month
README.md	Update README.md	last month
dataLoader.py	Add files via upload	2 months ago
loss.py	Add files via upload	last month
train.py	Add files via upload	last month

README.md

### A Light Weight Model for Active Speaker Detection

Ranked #4 Audio-Visual Active Speaker Detection on AVA-ActiveSpeaker

This repository contains the code and model weight for our paper (Accepted by CVPR 2023):

A Light Weight Model for Active Speaker Detection  
Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen

About

The repository for IEEE CVPR 2023 (A Light Weight Model for Active Speaker Detection)

Readme 32 stars 2 watching 3 forks Report repository

Releases

No releases published

Packages

No packages published

Languages

- Python 100.0%

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

# Thank you!



四川大學  
SICHUAN UNIVERSITY



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen