QuAVF: Quality-aware Audio-Visual Fusion for Ego4D Talking to Me Challenge

Hsi-Che Lin[1] Chien-Yi Wang[2] Min-Hung Chen[2] Szu-Wei Fu[2] Yu-Chiang Frank Wang[1,2]
[1] National Taiwan University [2] NVIDIA
hsichelin@gmail.com, {chienyiw, minhungc, szuweif, frankwang}@nvidia.com

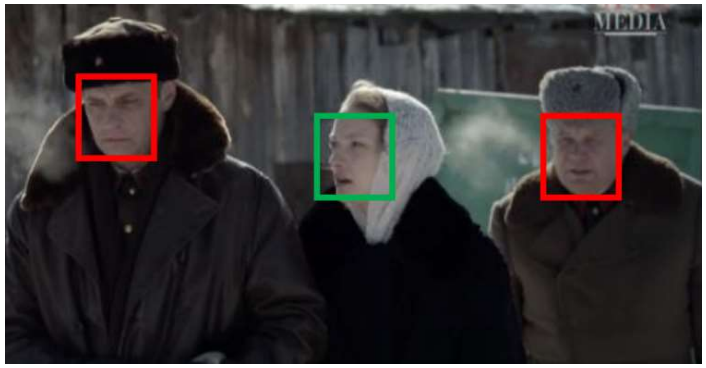
arxiv 2023

Background

⊡社交理解（人与人之间的互动）：

- 在由可穿戴相机拍摄的视频中，一个人与周围世界互动的沉浸式自我中心视角呈现了一个相互关联的视频网络理解任务——手部物体操作、空间导航或人与人之间的互动——这些任务在个人目标的驱动下不断展开。
- 以自我为中心的社交理解的进步可能会带来更强大的虚拟助理和社交机器人。

Background



Active Speaker Detection

left : Who is talking?

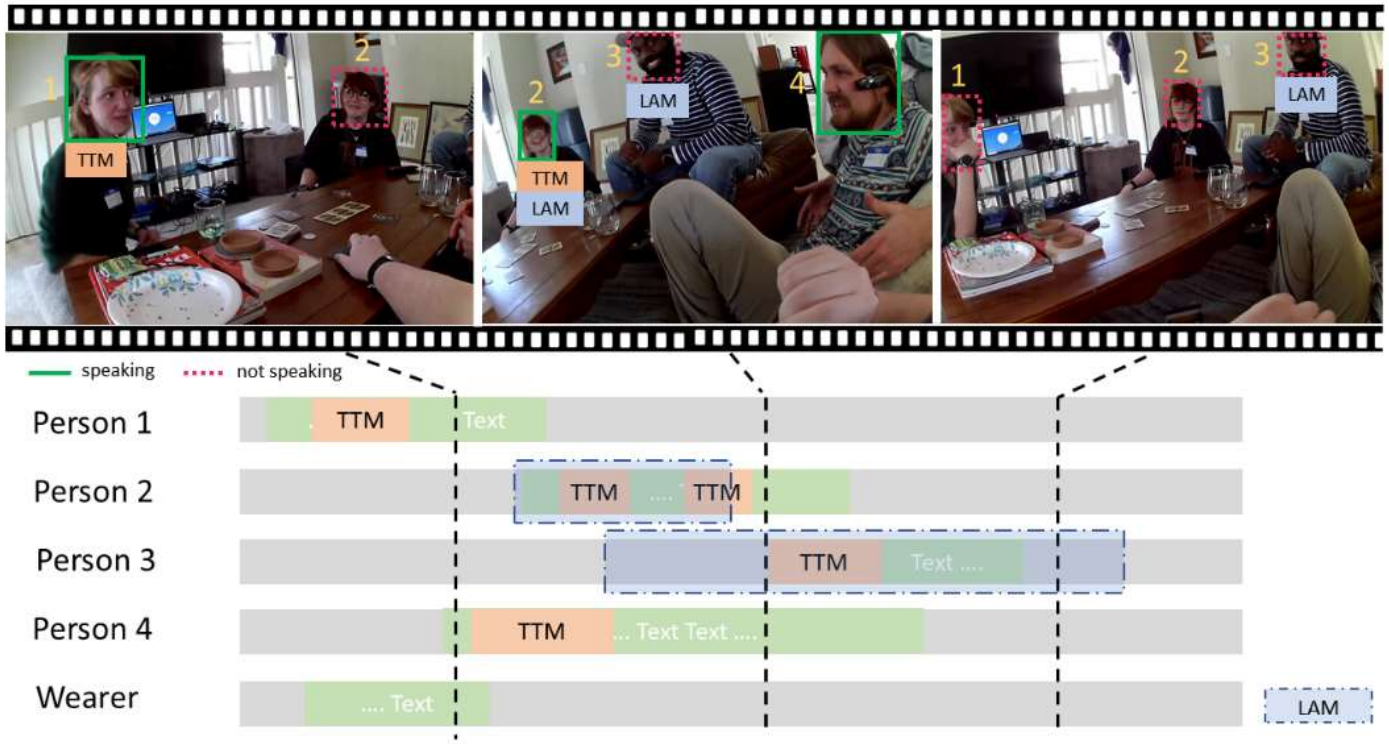
for more

Who is talking to “Me” (the camera-wearer) ?

Introduction

the Talking to Me (TTM) challenge :

- given a video and audio segment with the same tracked faces and an additional label that identifies speaker status, classify whether each visible face is talking to the camera wearer.



Introduction

Ego4D dataset:

- **74** worldwide locations and **9** countries, with over **3,670** hours of daily-life activity video.
- seven different head-mounted cameras were deployed across the dataset: GoPro, Vuzix Blade, Pupil Labs, ZShades, OR-DRO EP6, iVue Rincon 1080, and Weeview. They offer tradeoffs in the modalities available (RGB, stereo, gaze)
- Ego is for egocentric, and 4D is for 3D spatial plus temporal information.
- aims to catalyse the next era of research in first-person visual perception.

Method

baseline AV-joint Framework

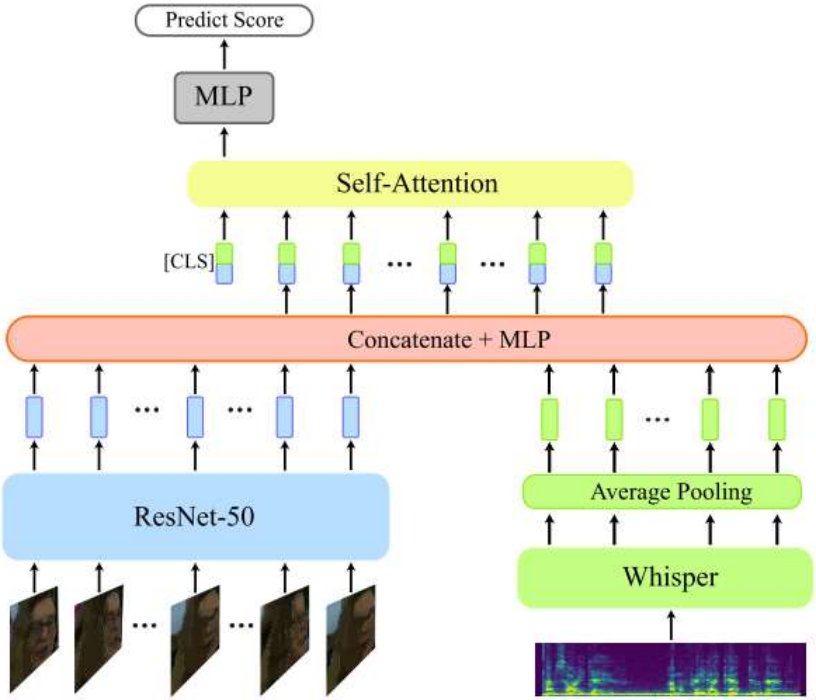


Figure 2. **The baseline AV-joint model approach.**

Q1:some training data lacks the corresponding bounding box label

- as described in the original Ego4D paper, the determination of the TTM label is based on vocal activity, irrespective of whether the person is visible in the scene.
- about 0.7M frames out of 1.7M frames with TTM label do not have bounding box label.

How to solve?

- discard data without bounding box labels?
- zero padding?

Method

baseline AV-joint Framework

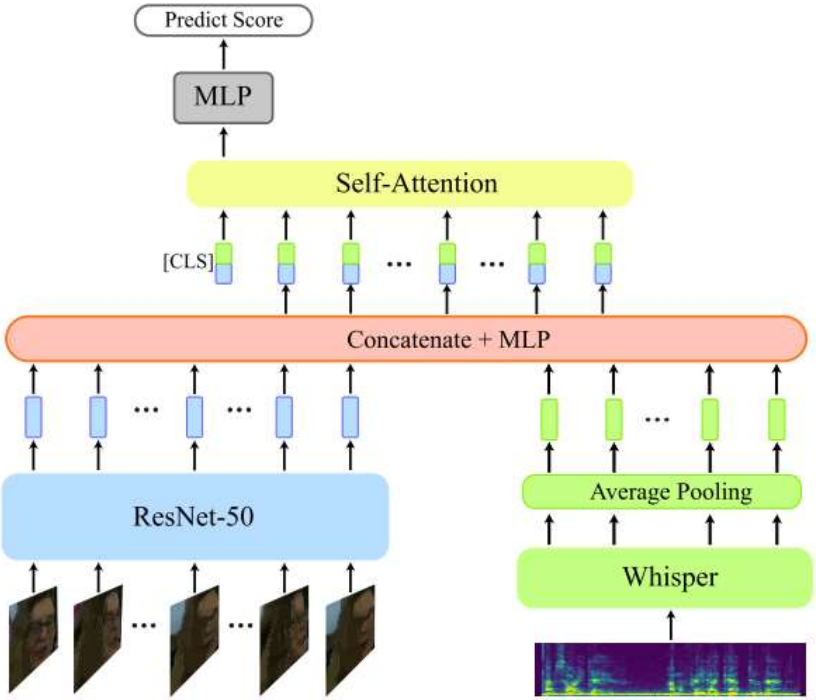


Figure 2. **The baseline AV-joint model approach.**

Q2:data quality

- limitations of the hardware used to record egocentric videos, and potential inaccuracies in bounding box annotations

How to solve?

Method

QuAVF Framework

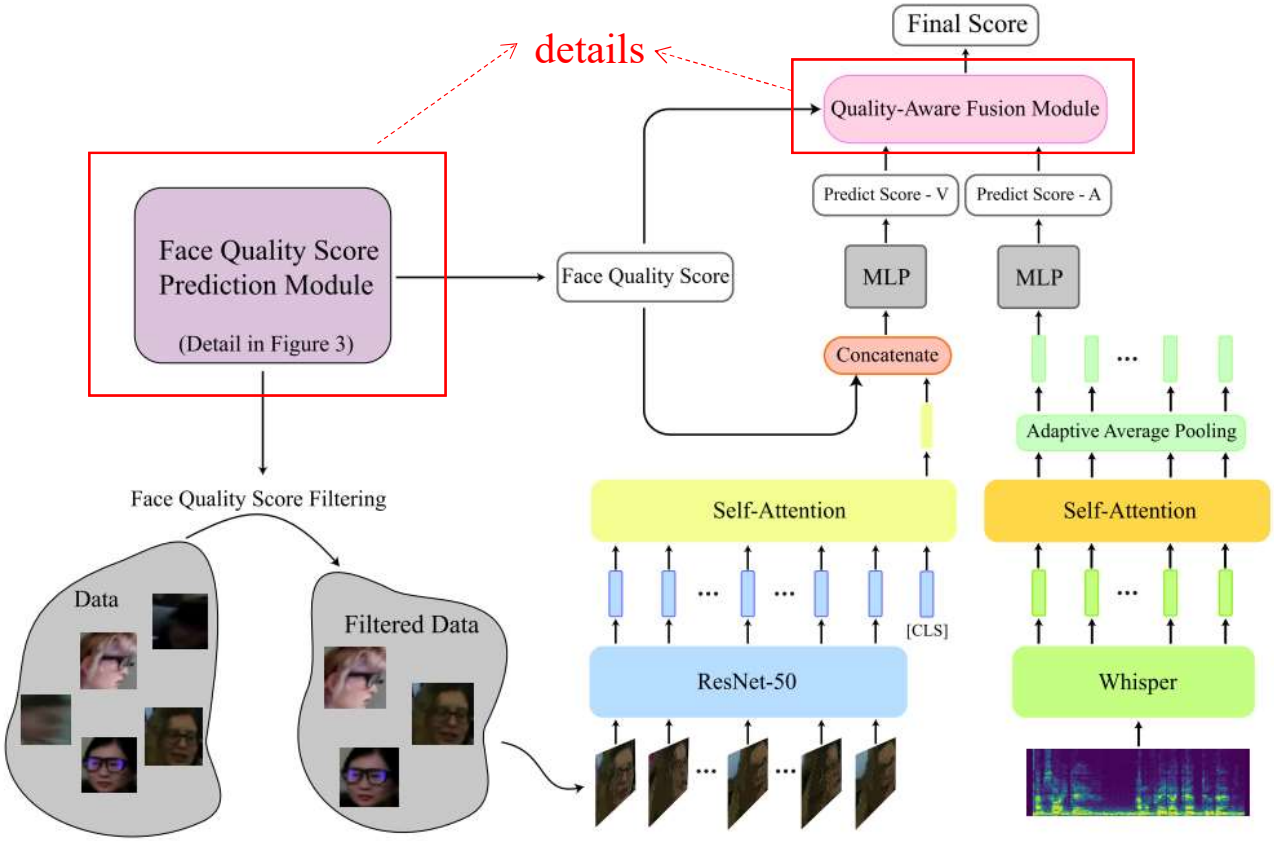


Figure 1. **An illustration of our proposed Quality-aware Audio-Visual Fusion (QuAVF) framework.**

audio:
fully utilize all the labels in the data-set, unaffected by variations in image quality

video:
take steps to ensure data quality by incorporating an additional model that provides a quality score indicating the likelihood of a face appearing in images.
This quality score is utilized to filter out inappropriate training data for the vision branch.

fusion:
Leveraging the same quality score, we introduce a quality-aware audio-visual fusion (QuAVF) approach.

Method

QuAVF Framework

- Face Quality Score Prediction Module

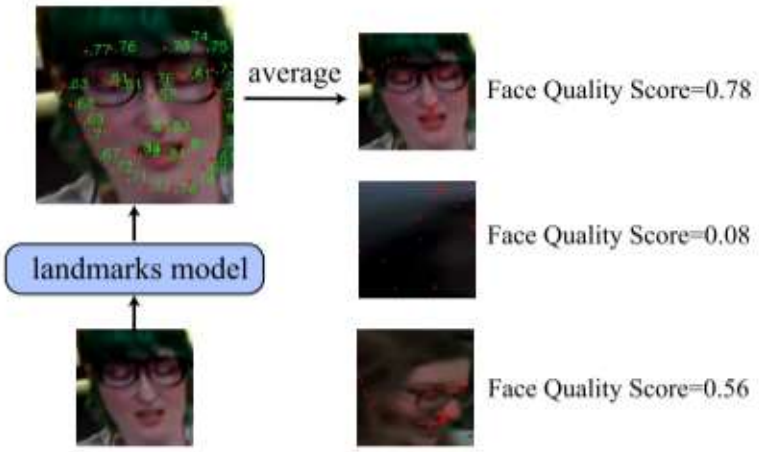


Figure 3. **An illustration of face quality score computation with the facial landmark model [1].**

apply the **facial landmarks prediction model [1]** on the bounding box region of training data and average the confidence scores of all the landmark points .
We treat the resulting score as the face quality score for that image, which represents how likely there is a face appearing in that region.

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE international conference on computer vision, pages 1021–1030, 2017. 3

Method

QuAVF Framework

- Quality-Aware Fusion Module

why need?

two independent model to process the audio and images separately

how?

considers the face quality score and fuse the prediction scores from two branches

compute the weighted sum of score from each branch with the weight of the vision branch set as the face quality score (the weight of the audio branch is then (1−face quality score))

$$score = (1 - facescore) \times a["score"] + facescore \times v["score"]$$

Experiments

Datasets and results:

- Ego4D dataset for soical interactions benchmark

Table 1. Results of Talking to Me (TTM) challenge.

method	Validation		Test	
	Accuracy	mAP	Accuracy	mAP
Random Guess [2]			47.41	50.16
ResNet-18 Bi-LSTM [2]			49.75	55.06
EgoTask Translation [6]			55.93	57.51
Baseline AV joint	58.1	59.5	55.66	57.05
Audio-only	71.7	70.1	**57.77**	**67.39**
Vision-only	64.0	67.2	54.80	56.17
QuAVF	**71.8**	**71.2**	57.05	65.83

mAP: mean average precision

[2] Ego4d: Around the world in 3,000 hours of egocentric video. CVPR 2022
[6] EgoT2 Egocentric Video Task Translation. CVPR 2023

Ablation Studies

Results of baseline AV joint model on validation data.

method	backbone	Accuracy	mAP
AV joint	ResNet-50 Whsiper	**53.6**	**59.5**
	AV-HuBERT [5]	53.4	58.2

Results of vision branch model on validation data

method	filter	face quality score	Accuracy	mAP
Vision branch			53	54
	> 0.5		51.3	55.9
	> 0.3		57.2	62
	> 0.3	scalar	57	62
	> 0.3	quantized	63	65
(fine-tune)	> 0.3	quantized	**64**	**67.2**

scalar:
apply a linear transformation on the scalar and concatenate the output feature with the final [CLS] token before the prediction head

quantized:(not give)
The result of quantization will be a onehot vector showing which level of magnitude the score falls into.
(面部质量分数属于哪个数量级)

[5] Learning audio-visual speech representation by masked multimodal cluster prediction. arxiv 2022

Discussion

Performance Gap between Validation and Test

Table 1. Results of Talking to Me (TTM) challenge.

method	Validation		Test	
	Accuracy	mAP	Accuracy	mAP
Random Guess [2]			47.41	50.16
ResNet-18 Bi-LSTM [2]			49.75	55.06
EgoTask Translation [6]			55.93	57.51
Baseline AV joint	58.1	59.5	55.66	57.05
Audio-only	71.7	70.1	**57.77**	**67.39**
Vision-only	64.0	67.2	54.80	56.17
QuAVF	**71.8**	**71.2**	57.05	65.83

Discussion

how to improve Visual branch?

eye contact ? (×)

head pose ? (×)

Thank you!