LoCoNet: *Lo*ng-Short *Co*ntext *Net*work for Active Speaker Detection

Xizi Wang[1] Feng Cheng[2] Gedas Bertasius[2]
[1]Indiana University [2]UNC Chapel Hill
xiziwang@iu.edu {fengchan,gedas}@cs.unc.edu

Abstract

Active Speaker Detection (ASD) aims to identify who is speaking in each frame of a video. Solving ASD involves using audio and visual information in two complementary contexts: long-term intra-speaker context models the temporal dependencies of the same speaker, while short-term inter-speaker context models the interactions of speakers in the same scene. Motivated by these observations, we propose LoCoNet, a simple but effective Long-Short Context Network that leverages Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM) in an interleaved manner. LIM employs self-attention for long-range temporal dependencies modeling and cross-attention for audio-visual interactions modeling. SIM incorporates convolutional blocks that capture local patterns for short-term inter-speaker context. Experiments show that LoCoNet achieves state-of-the-art performance on multiple datasets, with 95.2% (+0.3%) mAP on AVA-ActiveSpeaker, 97.2% (+2.7%) mAP on Talkies, and 68.4% (+7.7%) mAP on Ego4D. Moreover, in challenging cases where multiple speakers are present, LoCoNet outperforms previous state-of-the-art methods by 3.0% mAP on AVA-ActiveSpeaker. The code is available at https://github.com/SJTUwxz/LoCoNet_ASD.

1. Introduction

Real-world interactive computer vision systems need to recognize not only the physical properties of a scene, such as objects and people, but also the social properties, including how people interact with each other. One fundamental task is identifying, at any moment, who is speaking in a complex scene with multiple interacting individuals. This Active Speaker Detection (ASD) problem [2, 3, 11, 18, 24, 25, 29, 32, 35, 39, 40, 42, 53, 58, 59, 62, 65, 70] is crucial for many real-world applications like human-robot interaction [33, 61, 66], speech diarization [14, 16, 20, 21, 68, 73], video re-targeting [6, 44, 74], multimodal learning [4, 7, 17, 19, 28, 30, 34, 34, 46, 49, 50, 50, 54, 55, 55, 60, 63], etc.

How can we tell whether someone is speaking? Visual cues such as movements of the mouth and eyes, when correlated with audio signals, often offer direct and primary evidence. The inter-modality synchronization over a longer audio-visual segment also provides complementary information [2, 39, 48, 65]. The first row of Fig 2 shows how Long-term Intra-speaker Modeling helps discern this primary indicator by observing one person for a long time span.

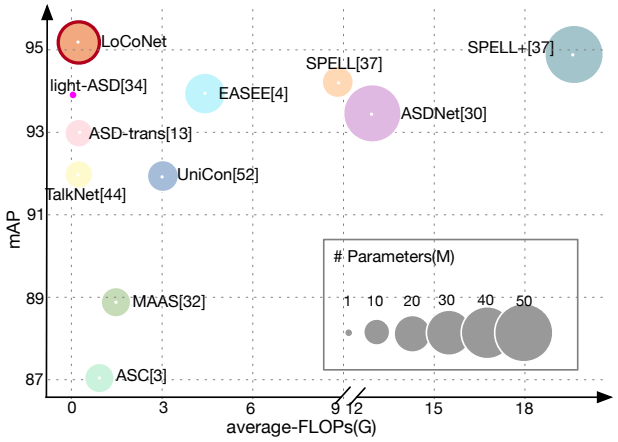

Figure 1. **Comparison of ASD methods in terms of mean average precision (mAP) on the AVA-ActiveSpeaker dataset, average-FLOPs, and number of parameters.** Note that average-FLOPs is the computation required to predict the speaking activity of one face crop.

However, in a complex video, a person's face is often occluded, turned away, off-frame, or very small, posing challenges for directly inferring speaking activity from visual cues. Fortunately, valuable evidence about the target speaker can be gleaned from the behaviors of others in the scene [61]. The second row of Fig 2 shows an example: from the left m-frame video segment, even with a partially visible face of the man on the right, it is easy to see that he is speaking at T_{i+m} because the woman in the middle turns her head ($T_{i+1} \rightarrow T_{i+m}$) and neither of the other two people open their mouths at T_{i+m}. Notably, the woman's gaze towards the man on the right at T_{i+m} does not provide substantial information on whether the man is speaking at a distant time T_{j+1}. Therefore, we argue that Inter-speaker Modeling is sufficient in short-term temporal windows, since activities

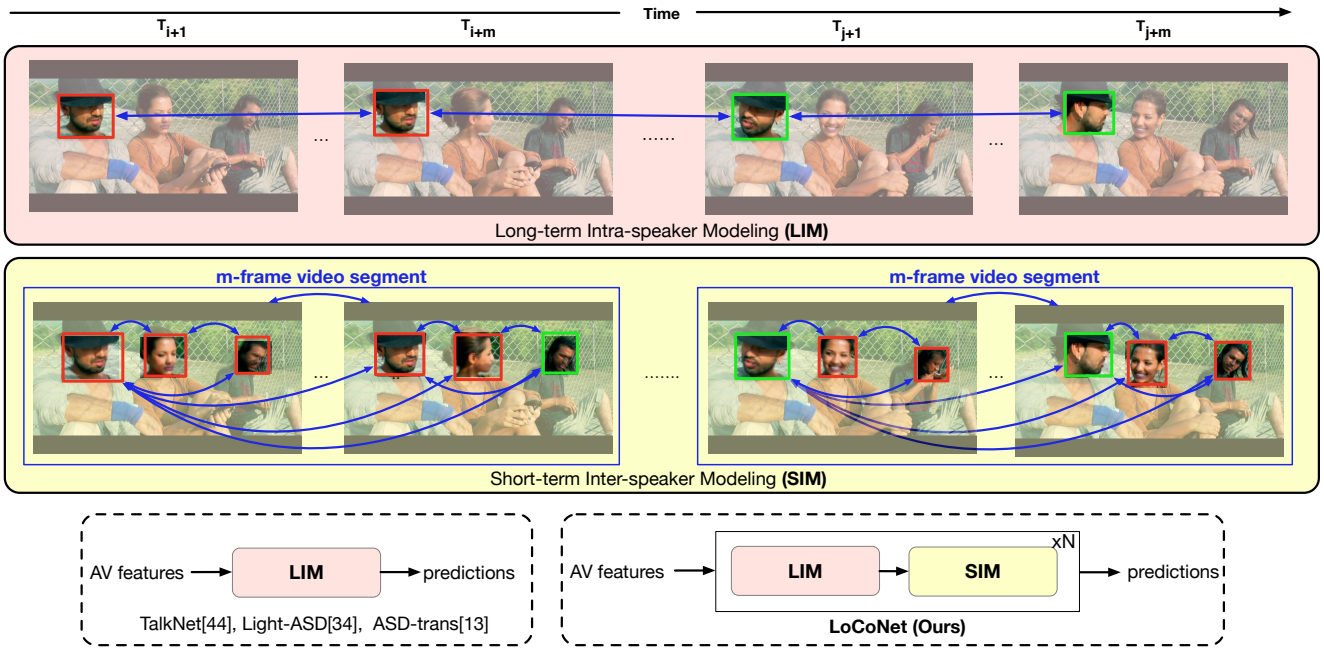

Figure 2. **Long-term Intra-speaker Modeling (LIM), Short-term Inter-speaker Modeling (SIM), and comparison of LoCoNet with existing long-term parallel-inference ASD methods**. Red boxes show inactive speakers and green boxes show active speakers. LIM uses the features of a single speaker across all frames to capture long-term relationships. SIM models the relationships of speakers within a short m-frame segment to capture the conversation pattern. The speaker context modeling of the existing long-term parallel-inference ASD methods [15, 45, 65] only focuses on LIM, while LoCoNet models LIM and SIM to learn both contexts.

of speakers within a short time range are more correlated than speakers separated farther away in time. Cognitive research [43, 57] also suggests that speaker-listener coupling is more coordinated in nearby frames.

To solve this per-frame video classification task, existing ASD methods can be put into two categories: 1) **parallel-inference** methods [15, 40, 45, 65, 75] take all frames as input and predict the results for all frames in one pass. 2) **sequential-inference** methods [2, 3, 39, 42, 48] take a short clip as input and only give prediction result of the center frame. Thus, a sliding window strategy is often adopted to get results for all the frames. As depicted in Fig. 1, the average-FLOPs of parallel-inference methods [15, 45, 65] to predict the speaking activity of one face crop are often much lower than sequential-inference methods [2, 3, 39, 42, 48]. However, most parallel-inference methods [15, 45, 65, 72] that take long video clips as input do not consider multi-speaker context, which could cause performance degradation as speakers' interaction is crucial for ASD task.

With the above issues in mind, we propose LoCoNet, an end-to-end Long-Short Context Network. Long-term Intra-speaker Modeling (LIM) employs a self-attention mechanism [67] for long-range dependencies modeling and a cross-attention mechanism for audio-visual interactions. Short-term Inter-speaker Modeling (SIM) incorporates convolutional blocks to capture local conversational patterns. Moreover, while most ASD methods use vision backbones for audio encoding [2, 3, 42, 65] due to high temporal down-sampling in most audio backbones [12, 23, 27, 38, 56], we propose VGGFrame to leverage pretrained AudioSet [22] weights to extract per-frame audio features. We also use a parallel inference strategy for more efficient video processing.

Our extensive experiments validate the effectiveness of our approach. On the AVA-ActiveSpeaker dataset [58], LoCoNet achieves 95.2% mAP, outperforming previous state-of-the-art method SPELL+ [48] by 0.3% with 38× less computational cost. Furthermore, LoCoNet achieves 97.2% (+2.7%) mAP on Talkies [42] and 68.4% (+7.7%) mAP on Ego4D's Audio-Visual benchmark [24]. LoCoNet works especially well in challenging scenarios such as with multiple speakers or small speaker faces.

2. Related Work

Most recent techniques for ASD can be characterized in terms of three salient dimensions: frame-level processing strategy that determines the inference speed of the method, the extracted context information to enhance the feature representations for prediction, and the training mechanisms.

Frame-level processing strategy. Given a long video, existing ASD methods employ two main strategies for generating per-frame predictions. 1) *Parallel-inference* [15, 40, 45, 65]

takes all frames of the video as input and predict the per-frame results in one pass. Such methods are often fast, as depicted in Fig. 1. However, they typically do not consider interactions among multiple speakers. 2) *Sequential-inference* [2, 3, 39, 42, 48] predicts one frame by sampling a short clip centered around that frame. They need a sliding window strategy to produce predictions for all frames, resulting in slower inference speed. Our proposed LoCoNet adopts a parallel inference strategy, combining fast inference speed with effective consideration of speakers' interactions.

Context Modeling. ASD benefits from both intra-speaker and inter-speaker contexts. TalkNet [65] models long-term temporal intra-speaker context to distinguish speaking and non-speaking frames. ASC [2] employs self-attention for long-term inter-/intra-speaker modeling, and an LSTM for long-term temporal refinement. ASDNet [39] aggregates short-term features of target speaker and background speakers at nearby frames, and leverages Bidirectional GRU [13] for long-term temporal modeling. Light-ASD [45] also uses Bidirectional GRU for temporal modeling. MAAS [42], EASEE [3], and SPELL [48] use Graph Convolutional Networks [37, 69] to model relationships between the visual nodes and audio nodes of context speakers. TS-TalkNet [31] explores the use of reference speech to assist ASD.

In our proposed LoCoNet, we introduce Long-Short Context Modeling (LSCM), composed of Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM) in an interleaved manner. LIM captures long-term temporal dependencies with self-attention, and audio-visual interactions with cross-attention. SIM incorporates inter-speaker convolutional blocks to learn speakers' interactions.

Training mechanisms. Training on long videos can be memory-intensive, prompting some prior work [2, 39, 42, 48] to adopt a multi-stage training mechanism. In this process, a short-term feature extractor is initially trained, and subsequently, a long-term context modeling network is trained on the features extracted by the pre-trained feature extractor. TalkNet [65], EASEE [3], UniCon [75], light-ASD [45], ASD-transformer [15], and ADENet [72] use end-to-end training, fully leveraging the learning capabilities of the model. Similarly, our proposed LoCoNet is also trained end-to-end, enabling joint optimization of audio-visual feature representation learning and context modeling.

3. LoCoNet

Given the stacked visual face track $V \in \mathbb{R}^{S\times T\times H\times W\times 1}$ and the audio Mel-spectrograms $A \in \mathbb{R}^{4T\times M}$, LoCoNet aims to predict the speaking activity $\hat{R} \in \mathbb{R}^{T}$ of the target person in each frame. S is the number of speakers including the target speaker and $S-1$ context speakers in the same scene. T is the temporal length of the face track. H and W are the height and width of each visual face crop. M is the number of frequency bins of the audio Mel-spectrograms.

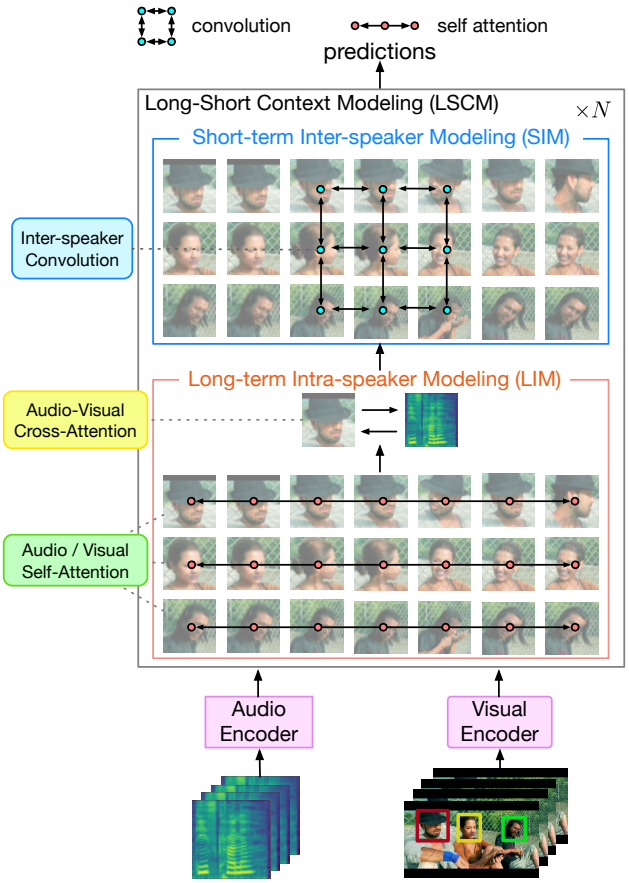

Figure 3. **An overview of LoCoNet.** Given a sequence of face tracks and audio of a target speaker, we sample $S-1$ speakers from all other people appearing in the scene and stack their face crops as visual input. Our method consists of 3 components: an audio encoder, a visual encoder, and a Long-Short Context Modeling module (LSCM) with N blocks, where each block includes an attention-based Long-term Intra-speaker Model (LIM) and a convolution-based Short-term Inter-speaker Model (SIM) for speaker interaction. LIM involves Audio-Visual Self-Attention for long-term intra-speaker dependencies and Audio-Visual Cross-Attention for audio-visual interaction. The final output is used to classify speaking activity of the target person across all frames.

As shown in Fig 3, LoCoNet consists of a visual encoder, an audio encoder, and a Long-Short Context Modeling (LSCM) module with N LSCM blocks. We explain each module in more detail below.

3.1. Encoders

Visual encoder. Given the face crop track $V_i \in \mathbb{R}^{T\times H\times W\times 1}$ of speaker v_i, the visual encoder yields a time sequence of visual embeddings $f_{v_i} \in \mathbb{R}^{T\times C}, i = 1, ..., S$. The stacked visual embeddings of the target speaker and all the sampled context speakers $f_v \in \mathbb{R}^{S\times T\times C}$ represent temporal context

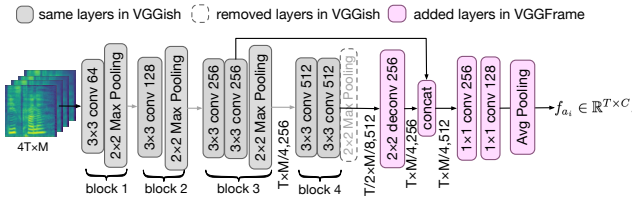

Figure 4. An illustration of our proposed audio encoder VGGFrame. We apply a deconvolutional layer to upsample the output feature of block-4. The output features of block-3 (before max pooling) and deconvolutional layer are concatenated and transformed to per-frame output features of shape $T \times C$.

of each speaker independently.

Audio Encoder. The audio encoder takes audio Mel-spectrograms $A \in \mathbb{R}^{4T\times M}$ as input. We need frame-level audio features for per-frame classification, but most pretrained audio encoders [12, 23, 27, 56] are for audio classification and thus have a high degree of temporal downsampling. To solve this problem, we propose VGGFrame as the audio encoder, which can fully utilize the pretrained VGGish [27]. The architecture of VGGFrame is illustrated in Fig 4. We remove the temporal downsampling layer after block-4 and add a deconvolutional layer to upsample the temporal dimension. We concatenate the intermediate features with the upsampled features to extract a hierarchy of representations. VGGFrame outputs the audio embeddings $f_{a_i} \in \mathbb{R}^{T\times C}$. To align with S speakers, we repeat f_{a_i} S times to produce the audio embedding $f_a \in \mathbb{R}^{S\times T\times C}$.

3.2. Long-Short Context Modeling

The visual and audio embeddings, derived independently by the visual and audio encoders, lack consideration of intra/inter-speaker context. Our Long-Short Context Modeling (LSCM) is designed to enhance these embeddings by learning long-term intra-speaker and short-term inter-speaker context in an interleaved manner. As shown in Fig. 3, LSCM consists of N blocks, each incorporating a Long-term Intra-speaker Modeling (LIM) module and a Short-term Inter-speaker Modeling (SIM) module consecutively. LIM limits the model to look at the same speaker across all frames, encouraging it to learn speaker-independent patterns from audio and visual interactions. In contrast, SIM constrains the model to examine all speakers in nearby frames and capture local interactions. Such inductive bias is instrumental in enhancing the model's capacity to glean valuable insights from these contextual dimensions.

LSCM inputs audio embeddings f_a and visual embeddings f_v, and produces context-aware embeddings $u_a^N, u_v^N \in \mathbb{R}^{S\times T\times C}$. These embeddings are concatenated to yield the final embeddings $u^N = concat(u_a^N, u_v^N)$. Via a linear layer, u^N is used to predict the speaking activities $\hat{R} \in \mathbb{R}^T$ of the target person. The computation process at each LSCM block l is detailed below.

3.2.1 Long-term Intra-speaker Modeling (LIM)

LIM consists of two submodules: (i) **Audio/Visual Self-Attention** models an individual person's behavior over a longer time period, and (ii) **Audio-Visual Cross-Attention** learns the interaction between audio and visual embeddings.

Audio/Visual Self-Attention. Since the model needs large capacity and to learn long-term dependencies, we employ the attention mechanism [67] with a Transformer layer applied on the temporal dimension to achieve long-term modeling,

$$\tilde{u}_v^l = \text{LN}(\text{MHA}(u_v^{l-1}, u_v^{l-1}, u_v^{l-1}) + u_v^{l-1}), \quad (1)$$
$$\tilde{u}_v^l = \text{LN}(\text{MLP}(\tilde{u}_v^l) + \tilde{u}_v^l), \quad (2)$$

where $u_v^0 = f_v$, $u_a^0 = f_a$, $u_v^{l-1} \in \mathbb{R}^{S\times T\times C}$ are the output visual embeddings of the previous LSCM block, $\text{LN}(\cdot)$ denotes Layer Normalization [5], $\text{MHA}(q, k, v)$ is multi-head attention [67] with query q, key k, value v, and MLP is a multi-layer perceptron. Audio Self-Attention is applied in the same way to u_a^{l-1} to obtain the audio output $\tilde{u}_a^l$.

Audio-Visual Cross-Attention. The visual and audio streams are so far processed separately. To enhance visual features with audio and vice versa, we use an Audio-Visual dual Cross-Attention,

$$\widehat{u}_v^l = \text{LN}(\text{MHA}(\tilde{u}_v^l, \tilde{u}_a^l, \tilde{u}_a^l) + \tilde{u}_v^l), \quad (3)$$
$$\widehat{u}_v^l = \text{LN}(\text{MLP}(\widehat{u}_v^l) + \widehat{u}_v^l), \quad (4)$$

where $\widehat{u}_v^l$ are audio-enhanced visual embeddings. Visual-enhanced audio embeddings $\widehat{u}_a^l$ are obtained in the same way via Audio-Visual Cross-Attention given $\tilde{u}_a^l$ and $\tilde{u}_v^l$.

3.2.2 Short-term Inter-speaker Modeling (SIM)

For a given moment in the video, the speaking activity of a target person is more coordinated with other speakers in closer frames [57], so the model should capture local temporal inter-speaker relationships. To do this, we employ a small **Inter-speaker Convolutional Network**,

$$u_v^l = \text{MLP}(\text{LN}(\text{Conv}_{s\times k}(\widehat{u}_v^l))) + \widehat{u}_v^l, \quad (5)$$
$$u_a^l = \text{MLP}(\text{LN}(\text{Conv}_{s\times k}(\widehat{u}_a^l))) + \widehat{u}_a^l, \quad (6)$$

where visual embeddings $u_v^l \in \mathbb{R}^{S\times T\times C}$ and audio embeddings $u_a^l \in \mathbb{R}^{S\times T\times C}$ are the output of the l-th LSCM block that will be passed to the next block. k is the temporal length of the receptive field and s is the number of speakers considered. Explicitly modeling inter-speaker context in nearby frames enables cross-frame inter-speaker information exchange. Our SIM module with a short temporal receptive field can help capture local dynamic patterns in interactions.

3.3. Training and Inference

Following [1, 10, 64], we train our model with multiple supervisions utilizing $u^N \in \mathbb{R}^{S\times T\times 2C}$ (Sec. 3.2) and u^i derived from each intermediate block of LSCM. For each u^i, a fully-connected (FC) layer is applied, yielding prediction results $\hat{R}^i \in \mathbb{R}^T$ corresponding to the target speaker for each frame. All FC layers share their parameters. The overall loss function is $L = \sum_{i=1}^{N} \text{CrossEntropy}(\hat{R}^i, R)$ where R is the ground-truth. During inference, we can reduce the computation by a factor of S by reusing speakers' features extracted by the visual encoder, which uses the most FLOPs.

3.4. Implementation details

Following [65], our visual encoder consists of a 3D convolutional layer, ResNet-18 [26], and a visual temporal convolution network (V-TCN) [41]. Our audio encoder is the proposed VGGFrame initialized with VGGish [12] weights pretrained on AudioSet [22]. We sample $S = 3$ speakers and $T = 200$ frames. In SIM, s is set to 3, which is the same as S, and k is set to 7 frames. The face crops are resized to 112×112. Visual augmentation includes randomly resized cropping, horizontal flipping, and rotation. For audio augmentation, another audio signal is randomly chosen from the rest of the training set and added as noise to the target audio. We train LoCoNet with Adam [36] for 25 epochs on 4 RTX6000 GPUs with batch size 4 using PyTorch [51]. The learning rate is 5×10^{-5} and reduced by 5% each epoch.

4. Experimental Setup

4.1. Datasets

AVA-ActiveSpeaker [58] is a standard benchmark for ASD, consisting of 262 videos from Hollywood movies with 3.65 million frames and 5.3 million face crops. Following [3, 15, 31, 45, 47, 48], we evaluate on the validation set.

Talkies [42] is an in-the-wild ASD dataset with 23,507 face tracks extracted from 421,997 labeled frames. It focuses on challenging cases with more speakers, diverse actors and scenes, and more off-screen speech.

Ego4D [24]'s Audio-Visual benchmark has 3,670 hours of egocentric video of unscripted daily activities in many environments. It includes many challenges that complement exocentric benchmarks, including unusual viewpoints and speakers who are off-screen.

4.2. Evaluation Metric

Following [2, 3, 39, 42, 65], we use the official ActivityNet [9] evaluation tool to compute mean average precision (mAP) and evaluate on the AVA-ActiveSpeaker [58]. We also compute AUC [8] as another evaluation metric using Sklearn [52]. We use mAP to evaluate on Talkies [42] and Ego4D [24].

Method	Vid. Enc.	Params	*average*-FLOPs	mAP	AUC
Multi-Stage					
ASC [2]	R-18 [26]	$23.0M$	$1.0G$	87.1	-
MAAS [42]	R-18	$23.0M$	$1.6G$	88.8	-
ASDNet [39]	3DRNext-101	$51.0M^\dagger$	$13.2G^\dagger$	93.5	-
SPELL [48]	R-18+TSM	$23.5M^\dagger$	$8.7G^\dagger$	94.2	-
SPELL+ [48]	R-50+TSM	$51.23M^\dagger$	$19.6G^\dagger$	94.9	-
End-to-End					
UniCon [75]	R-18	$23.8M$	$3.0G^\dagger$	92.2	97.0
TalkNet [65]	R-18+VTCN	$15.7M$	$0.51G$	92.3	96.8
ASD-Trans [15]	R-18+VTCN	$15.0M$	$0.55G$	93.0	-
EASEE [3]	3D R-50	$26.8M^\dagger$	$4.3G^\dagger$	94.1	-
Light-ASD [45]	(2+1)D Conv	$\mathbf{1.02M}$	$\mathbf{0.2G}$	94.1	-
TS-TalkNet [31]	R-18+VTCN	$36.8M$	$2.3G$	93.9	-
LoCoNet(Ours)	R-18+VTCN	$34.3M$	$0.51G$	**95.2**	**98.0**

Table 1. **Comparison with SOTAs on AVA-ActiveSpeaker**. 3DRNext denotes 3D ResNext [71]. R denotes 2D ResNet [26]. *average*-FLOPs represents the averaged FLOPs needed to process a single face crop. † denotes our estimates based on their visual encoders. Most methods incur higher costs by extracting features for each frame through stacking multiple adjacent frames (*i.e.*, 11 in SPELL). LoCoNet achieves the highest mAP with modest FLOPs.

5. Results and Analysis

We first compare the proposed method LoCoNet with previous state-of-the-art methods on multiple datasets and challenging scenarios. Then we validate our hypotheses of long-term intra-speaker and short-term inter-speaker modeling. Finally, following [2, 39, 45, 65, 75], we conduct extensive ablations on each component of LoCoNet (on AVA-ActiveSpeaker [58] unless otherwise noted).

5.1. Comparison with State-of-the-Art

In this section, we compare our approach with state-of-the-art methods on the three datasets.

AVA-ActiveSpeaker. From Table 1, end-to-end methods exhibit fewer FLOPs while maintaining competitive mAP compared to multi-stage methods. The higher FLOPs of multi-stage methods stem from their sequential-inference processing strategy (Sec. 2) of stacking multiple neighboring frames centered at time t. LoCoNet achieves a 95.2% mAP, surpassing the best-performing end-to-end ASD method Light-ASD [45] by 1.1% while using modest average-FLOPs. Moreover, LoCoNet outperforms previous state-of-the-art multi-stage method SPELL+ [48] by 0.3% with about 32% fewer parameters and over 38× fewer average-FLOPs.

Talkies set. We evaluate LoCoNet with other methods under three training settings: (i) AVA-ActiveSpeaker, (ii) Talkies, and (iii) pretrained on AVA-ActiveSpeaker and finetuned on Talkies. As shown in Table 2, LoCoNet outperforms EASEE by 1.7%, 2.5%, and 2.7% in these three settings, respectively.

Method	Train Set		mAP
	AVA	Talkies	(%)
MAAS [42]	✓	✗	79.7
EASEE [3]	✓	✗	86.7
LoCoNet	✓	✗	**88.4**
EASEE [3]	✗	✓	93.6
light-ASD[45]	✗	✓	93.9
LoCoNet	✗	✓	**96.1**
EASEE [3]	✓	✓	94.5
LoCoNet	✓	✓	**97.2**

Table 2. **Comparison on Talkies dataset** under three training settings: train on AVA-ActiveSpeaker alone, train on Talkies alone, or train on AVA-ActiveSpeaker and finetune on Talkies.

Method	mAP (%)
TalkNet [65]	51.7
Challenge Winner [47]	60.7
LoCoNet	**68.4**

Table 3. **Comparison on Ego4D dataset.** The challenge winner is not specifically optimized for ASD but the large improvement (+7.7%) still shows the strong generalizability of LoCoNet. The result of TalkNet was obtained by us with their released code.

It also outperforms light-ASD by 2.2% when both models are trained on Talkies only.

Ego4D dataset. We evaluate our method on Ego4D Audio-Visual benchmark [24]. LoCoNet achieves 68.4% mAP, outperforming TalkNet and Challenge Winner [47] (SPELL [48]-based) by 16.7% and 7.7% respectively. Egocentric videos, characterized by constant camera motion, lower clarity, and more complex scenes compared to exocentric videos, demonstrate the potential of the proposed Long-Short Context Modeling in real-life scenarios. Our approach benefits from capturing both multi-speaker interaction and single-speaker behavior.

5.2. Challenging Scenario Evaluation

Quantitative analysis. We report the performance of LoCoNet on AVA-ActiveSpeaker under different face sizes: (i) Small: faces with width less than 64 pixels; (ii) Medium: faces with width between 64 and 128 pixels; (iii) Large: faces with width larger than 128 pixels. We also study the effect of the number of visible faces in a video frame (1, 2, or 3).

The results, along with the portions of each category, are shown in Tables 4 and 5 respectively. LoCoNet consistently performs the best across all scenarios, exhibiting the most significant improvement in the challenging multi-speaker case: +3.0% for 3 faces. This suggests that our method more effectively models both the target speaker's speaking pattern and the interactions of context speakers, allowing accurate inference of the speaking activity of the target person.

Method	Face Size		
	Small (18%)	Medium (30%)	Large (52%)
ASC [2]	56.2	79.0	92.2
MAAS [42]	55.2	79.4	93.0
TalkNet [65]	63.7	85.9	95.3
ASDNet [39]	74.3	89.8	96.3
EASEE [3]	75.9	90.6	96.7
light-ASD [45]	77.5	91.2	96.5
LoCoNet	**77.8**	**93.0**	**97.3**

Table 4. **Results as a function of face size.** LoCoNet achieves the highest mAP among all face sizes.

Method	# Faces		
	1 (45%)	2 (33%)	3 (11%)
ASC [2]	91.8	83.8	67.6
MAAS [42]	93.3	85.8	68.2
TalkNet [65]	95.4	89.6	80.3
ASDNet [39]	95.7	92.4	83.7
EASEE [3]	96.5	92.4	83.9
light-ASD [45]	96.2	92.6	84.4
LoCoNet	**97.0**	**94.6**	**87.4**

Table 5. **Results as a function of visible faces in the scene.** Larger improvements are observed on more challenging cases (*i.e.*, 3 faces).

Qualitative analysis. Fig 5 visualizes the results of LoCoNet and TalkNet [65] on AVA-ActiveSpeaker [58] with the groundtruth labels. The video on the left shows four visible speakers talking, posing a challenge in distinguishing the active speaker amid multiple discussions in the same scene. LoCoNet accurately locates the active speakers in this case, whereas TalkNet fails to recognize some of them. The first two columns of the video on the right shows a woman with a very small visible face as the active speaker, while the man with a large visible face is not speaking. TalkNet fails to locate the active speaker while LoCoNet succeeds. By combining long-term intra-speaker context to compare the speaking pattern of each individual and short-term inter-speaker context to examine the conversations, our approach better overcomes this challenging speaking scenarios. However, in the last column, both methods fail to recognize the active speaker at the back. This scenario is especially challenging as the two active speakers are in separate conversations with one being far less salient than the other, making it difficult to infer the speaking activity of the less salient speaker.

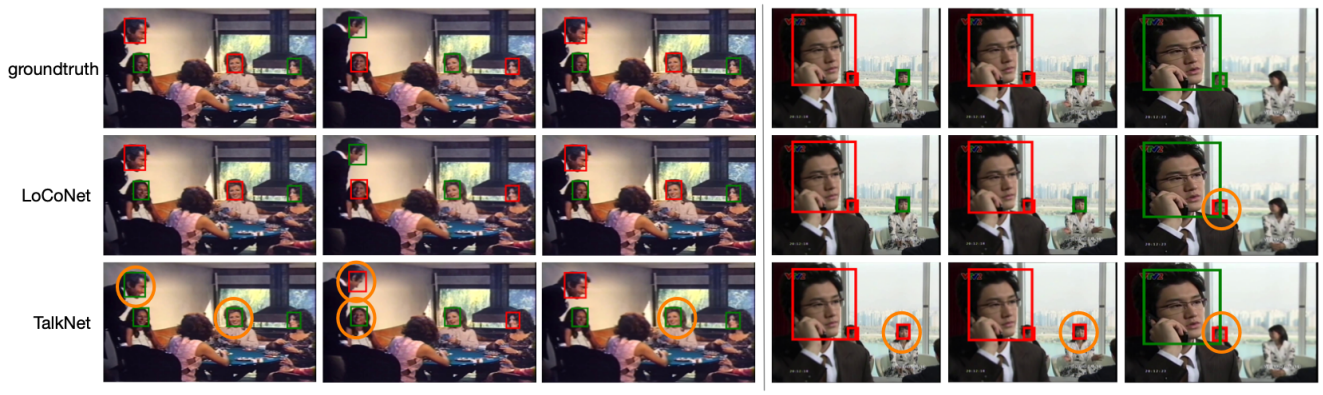

Figure 5. **Results comparison of LoCoNet and TalkNet on challenging scenarios of AVA-ActiveSpeaker.** Red boxes denote not-active speaker. Green box denote active speaker. Orange circles refer to false predictions. The video on the left shows a multi-people conversation with four speakers, and separate conversation of two. The video on the right shows an active speaker with a small face. Both scenes are challenging, and LoCoNet predicts accurately in most cases.

5.3. Attention visualizations of LIM and SIM

We next visualize the effectiveness of Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM). The left part of Fig. 6 visualizes the attention weights across different frames of a single speaker in LIM. It is evident that speaking and non-speaking activities are distinctly separated, with clear boundaries when speaking activities change. This verifies that LIM contributes to accurate speaking activity detection. The right part of Fig. 6 shows the portions of information (measured by L2-norm, as convolutions are used for SIM) drawn by the target speaker at target frame from all speakers at nearby frames (and information flows with very small portions are not shown). The woman on the left gradually turns her face to the target speaker, which is the most indicative sign that the target speaker has started speaking. The distribution of information flow reveals that SIM infers from the behaviors of context speakers, assigning more attention to the woman on the left.

5.4. Context Modeling Analysis

Does Intra-speaker Modeling require long-term? We train LoCoNet by keeping the number of speakers S as 1, and varying the temporal length of input frames T from 20 to 400. Table 6 indicates that the network performs worst when trained with the shortest video segments of 20 frames (0.8 sec). Performance improves as video segments become longer, with a 5.0% mAP increase at 100 frames (4 sec) and an additional 0.4% at 200 frames. This underscores the importance of long-term temporal context in intra-speaker modeling, aligning with findings from TalkNet [65] and ASC [2] that long-term temporal context provides better evidence of a speaking episode. We found 200 frames to be a good balance between performance and memory cost.

Is Short-term Inter-speaker Modeling Sufficient? We first validate the importance of inter-speaker context. Keeping the temporal length at 200 frames, we vary the number of speakers S from 1 to 3. Larger S values were not considered because $>$ 99% of videos in AVA-ActiveSpeaker [58] have at most 3 speakers in the same scene. Table 7 demonstrates that performance increases with more speakers included in

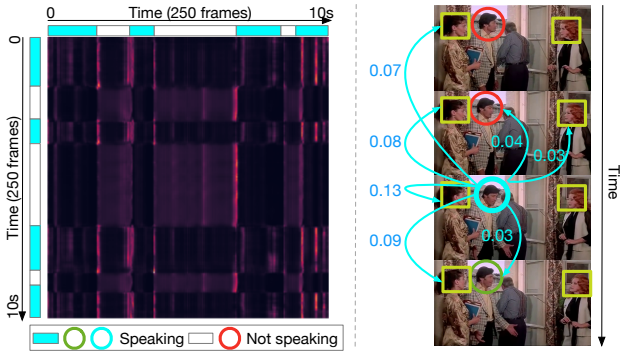

Figure 6. **Visualization of LIM and SIM.** Left: attention weights of one speaker in LIM, showing the ability to capture long-term intra-speaker context. Right: to predict the speaking activity of the target speaker at target frame, we show the information drawn from not-speaking target and speaking target and context speakers at nearby frames.

# Frames	20	100	200	300	400
mAP (%)	87.8	92.8	**93.2**	93.2	OOM

Table 6. **Temporal Length** in Long-term Intra-speaker Modeling. We set $S = 1$ so the context is intra-speaker only.

Speakers	1	2	3
mAP (%)	93.2	94.6	**95.2**

Table 7. **Number of Speakers** S in Short-term Inter-speaker Modeling. We set the number of frames T to 200.

Receptive Field	1	7	15	31	101
GFLOPs	**0.28**	0.98	1.93	3.82	12.1
mAP(%)	94.5	**95.2**	**95.2**	95.0	94.9

Table 8. **Temporal Receptive Field** k in SIM, indicating how many neighboring frames each speaker can explicitly query on in the SIM module. S is set as 3 and T as 200.

Ablation Settings				mAP
R-34	VGGFrame	LIM	SIM	(%)
✓				91.2
	✓			92.8
	✓	✓		94.1
	✓		✓	94.0
	✓	✓	✓	**95.2**

Table 9. **Ablations on components of LoCoNet.** We study the efficacy of audio encoder (ResNet-34 or VGGFrame), Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM). LIM and SIM are stacked 3 times if applied.

training. This supports our hypothesis that modeling multiple speakers in the inter-speaker context is necessary for ASD.

Next, with $S = 3$, we vary the temporal receptive field k of SIM from 1 (40 msec) to 101 (4 sec). Table 8 shows a 0.7% performance increase when increasing k from 1 to 7, confirming our assumption that motion in short-term inter-speaker context is more valuable than in-frame context. Performance saturates with further increases in the receptive field, and the computation cost increases drastically. This reinforces our hypothesis in Sec.1 that short-term inter-speaker modeling is sufficient.

5.5. Other Ablations

Does each component of LoCoNet help? In Table 9, replacing ResNet-34 [26] with VGGFrame as the audio encoder enhances mAP by 1.6%, showing the effectiveness of an audio encoder pretrained on audio datasets compared to a common vision encoder. With VGGFrame as audio encoder, adding 3 Long-term Intra-speaker Modeling (LIM) modules increases the performance by 1.3%, while 3 Short-term Inter-speaker Modeling (SIM) modules adds 1.2%. Both together improve mAP by 2.4%.

Convolution versus Window Self-Attention in SIM. Besides Convolution, we also try Window Self-Attention to capture local patterns for SIM (Eqn. 5). In Table 10, Convolution outperforms Window Self-Attention by 0.8% mAP, highlighting the superiority of Convolution in modeling local patterns of speakers' interaction.

Number of blocks in LSCM. We vary the number of blocks N in Long-Short Context Modeling (Sec. 3.2). Adding 1 block of LSCM yields a performance gain of 1.5%, showing the effectiveness of LIM and SIM. Adding two more blocks further increases by 0.7%, but results saturate at $N = 3$.

Design	mAP(%)
Convolution	**95.2**
Window Self-Attention	94.4

Table 10. **Ablation on designs of Short-term Inter-speaker Modeling**. We implement SIM using Convolution or Window Self-Attention (Eqn. 5). We keep the reception field the same (*i.e.*, 7 frames) and stack it three times ($N = 3$).

N	0	1	2	3	4	5
mAP(%)	92.8	94.3	95.0	**95.2**	95.0	95.0

Table 11. **Ablation on the number of blocks in LSCM** (N)**.** $N = 0$ refers to LoCoNet with no context modeling.

6. Conclusion

In this work, we observe that speaker activity can be more efficiently inferred from long-term intra-speaker context and short-term inter-speaker context. We thus design an end-to-end long-short context ASD framework that uses self-attention and cross-attention mechanisms to model long-term intra-speaker context and a convolutional network to model short-term inter-speaker context. With a simple backbone network, our method achieves state-of-the-art performance on 3 mainstream ASD benchmarks and significantly outperforms previous state of the art methods by 7.7% on Ego4D. We also show that in challenging scenarios where multiple speakers are in the same scene or speakers have small faces, our proposed method also outperforms previous methods. All of these results show the robustness and effectiveness of our method. Similar to existing long-term ASD methods, our method utilizes 8 sec of context. Future work should study how to implement larger contexts. Additionally, enhancing ASD in egocentric datasets could include adding other modalities, such as eye gaze.

Acknowledgements

The authors gratefully acknowledge Prof. David Crandall's guidance and feedback on earlier versions of this work. This work was supported in part by the National Science Foundation under award DRL-2112635 to the AI Institute for Engaged Learning, Sony Faculty Innovation Award, Laboratory for Analytic Sciences via NC State University, and ONR Award N00014-23-1-2356. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3159–3166, 2019. 5

[2] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 1, 2, 3, 5, 6, 7

[3] Juan Leon Alcazar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. *arXiv preprint arXiv:2203.14250*, 2022. 1, 2, 3, 5, 6

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[6] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 1

[7] Cigdem Beyan, Muhammad Shahid, and Vittorio Murino. Realvad: A real-world dataset and a method for voice activity detection by body motion analysis. *IEEE Transactions on Multimedia*, 23:2071–2085, 2020. 1

[8] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. 5

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 5

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[11] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *European Conference on Computer Vision*, pages 285–301. Springer, 2016. 1

[12] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 4, 5

[13] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 3

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

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

[16] Yifan Ding, Yong Xu, Shi-Xiong Zhang, Yahuan Cong, and Liqiang Wang. Self-supervised learning for audio-visual speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4367–4371. IEEE, 2020. 1

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

[18] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009. 1

[19] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1

[20] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934. IEEE, 2017. 1

[21] Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1086–1099, 2017. 1

[22] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2, 5

[23] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 2, 4

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 5, 6

[25] Ilya Gurvich, Ido Leichter, Dharmendar Reddy Palle, Yossi Asher, Alon Vinnikov, Igor Abramovski, Vishak Gopal, Ross Cutler, and Eyal Krupka. A real-time active speaker detection system integrating an audio-visual signal with a spatial querying mechanism. *arXiv preprint arXiv:2309.08295*, 2023. 1

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 8

[27] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2, 4

[28] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9844–9859, 2021. 1

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

[30] Arindam Jati and Panayiotis Georgiou. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1577–1589, 2019. 1

[31] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li. Target active speaker detection with audio-visual cues. *arXiv preprint arXiv:2305.12831*, 2023. 3, 5

[32] Chaeyoung Jung, Suyeon Lee, Kihyun Nam, Kyeongha Rho, You Jin Kim, Youngjoon Jang, and Joon Son Chung. Talknce: Improving active speaker detection with talk-aware contrastive learning. *arXiv preprint arXiv:2309.12306*, 2023. 1

[33] Soo-Han Kang and Ji-Hyeong Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, pages 1–11, 2021. 1

[34] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1

[35] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Look who's talking: Active speaker detection in the wild. *arXiv preprint arXiv:2108.07640*, 2021. 1

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

[37] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[38] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 2

[39] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 1, 2, 3, 5, 6

[40] Minyoung Kyoung and Hwa Jeon Song. Modeling long-term multimodal representations for active speaker detection with spatio-positional encoder. *IEEE Access*, 2023. 1, 2

[41] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016. 5

[42] Juan León-Alcázar, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. *arXiv preprint arXiv:2101.03682*, 2021. 1, 2, 3, 5, 6

[43] Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015. 2

[44] Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, and C-C Jay Kuo. Depth-aware stereo video retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2018. 1

[45] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023. 2, 3, 5, 6

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

[47] Kyle Min. Intel labs at ego4d challenge 2022: A better baseline for audio-visual diarization. *arXiv preprint arXiv:2210.07764*, 2022. 5, 6

[48] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 371–387. Springer, 2022. 1, 2, 3, 5, 6

[49] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 1

[50] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 1

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5

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

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

[55] Varshanth R Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Decompose the sounds and pixels, recompose the events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2144–2152, 2022. 1

[56] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018. 2, 4

[57] Daniel Richardson, Rick Dale, and Kevin Shockley. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. *Embodied communication in humans and machines*, pages 75–94, 2008. 2, 4

[58] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1, 2, 5, 6, 7

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

[60] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. S-vvad: Visual voice activity detection by motion segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2332–2341, 2021. 1

[61] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178, 2021. 1

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

[63] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017. 1

[64] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 5

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

[66] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223, 2016. 1

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

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

[69] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 3

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

[71] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

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

[73] Eric Zhongcong Xu, Zeyang Song, Chao Feng, Mang Ye, and Mike Zheng Shou. Ava-avd: Audio-visual speaker diarization in the wild. *arXiv preprint arXiv:2111.14448*, 2021. 1

[74] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5306–5315, 2020. 1

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