End-to-End Active Speaker Detection

Juan León Alcázar[1], Moritz Cordes[1,2], Chen Zhao[1], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology, KAUST
jc.leon@uniandes.edu.co, {chen.zhao,bernard.ghanem}@kaust.edu.sa
[2] Leuphana University Lüneburg moritz.cordes@stud.leuphana.de

Abstract. Recent advances in the Active Speaker Detection (ASD) problem build upon a two-stage process: feature extraction and spatio-temporal context aggregation. In this paper, we propose an end-to-end ASD workflow where feature learning and contextual predictions are jointly learned. Our end-to-end trainable network simultaneously learns multi-modal embeddings and aggregates spatio-temporal context. This results in more suitable feature representations and improved performance in the ASD task. We also introduce interleaved graph neural network (iGNN) blocks, which split the message passing according to the main sources of context in the ASD problem. Experiments show that the aggregated features from the iGNN blocks are more suitable for ASD, resulting in state-of-the art performance. Finally, we design a weakly-supervised strategy, which demonstrates that the ASD problem can also be approached by utilizing audiovisual data but relying exclusively on audio annotations. We achieve this by modelling the direct relationship between the audio signal and the possible sound sources (speakers), as well as introducing a contrastive loss.

1 Introduction

In active speaker detection (ASD), the current speaker must be identified from a set of available candidates, which are usually defined by face tracklets assembled from temporally linked face detections [35,5,28]. Initial approaches to the ASD problem focused on the analysis of individual visual tracklets and the associated audio track, aiming to maximize the agreement between the audio signal and the visual patterns [35,9,49]. Such an approach is suitable for scenarios where a single visual track is available. However, in the general (multi-speaker) scenario, this naive correspondence will suffer from false positive detections, leading to incorrect speech-to-speaker assignments.

Current approaches for ASD rely on two-stage models [28,25,40]. First, they learn to associate the facial motion patterns and its concurrent audio stream by optimizing a multi-modal encoder [35]. Then, this encoder serves as a feature extractor for a second stage, in which multi-modal embeddings from multiple speakers are fused [1]. These two-stage approaches are currently preferred given the technical challenges of end-to-end training with video data. Despite the computational efficiency of these approaches, their two-stage nature precludes them

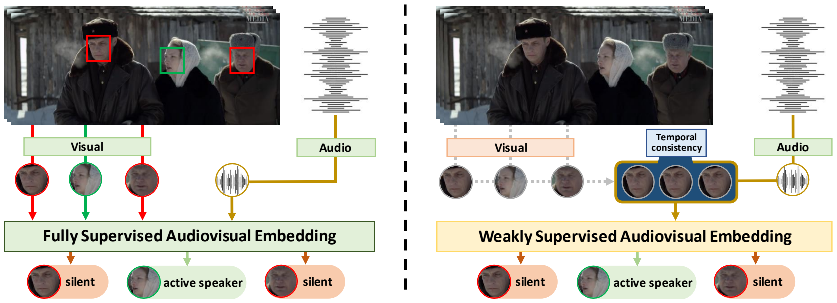

Fig. 1: **Fully and weakly-supervised audiovisual embeddings.** In the fully supervised scenario (left), we use the face crops as visual data and the Mel-frequency cepstral coefficients as audio data, we rely on visual and audio labels to directly optimize a shared feature embedding. In contrast, in the weakly supervised scenario, we omit the visual labels and optimize using only audio supervision. By modeling the visual-temporal consistency and speech-to-speaker assignments, we are able to optimize a shared embedding that can detect the active speakers without any visual supervision.

from fully leveraging the learning capabilities of modern neural architectures, namely directly optimizing the features for the multi-speaker ASD task.

In this paper, we present a novel alternative to the traditional two-stage ASD methods, called End-to-end Active Speaker dEtEction (EASEE), which is the first end-to-end pipeline for active speaker detection. Unlike conventional methods, EASEE is able to learn multi-modal features from multiple visual tracklets, while simultaneously modeling their spatio-temporal relations in an end-to-end manner. As a consequence, EASEE feature embeddings are optimized to capture information from multiple speakers and enable effective speech-to-speaker assignments in a fully supervised manner. To generate its final predictions, our end-to-end architecture relies on a spatio-temporal module for context aggregation. We propose an interleaved Graph Neural Network (iGNN) block to model the relationships between speakers in adjacent timestamps. Instead of greedily fusing all available feature representations from multiple timestamps, the iGNN block provides a more principled way of modeling spatial and temporal interactions. iGNN performs two message passing steps: first a spatial message passing that models local interactions between speakers visible at the same timestamp, and then a temporal message passing that effectively aggregates long-term temporal information.

Finally, EASEE's end-to-end nature allows the use of alternative supervision targets. In this paper, we propose a weakly-supervised strategy for ASD, named EASEE-W (shown in Figure 1). EASEE-W relies exclusively on audio labels, which are easier to obtain, to train the whole architecture. To optimize our network without the visual labels, we model the inherent structure in the ASD task, namely the direct relationship between the audio signal and its possible sound sources, *i.e.* , the speakers.

Contributions. This paper proposes EASEE, a novel approach for active speaker detection. Its end-to-end nature enables direct optimization of audio-visual embeddings and leverages novel training strategies, namely weak supervision. Our work brings the following contributions: (1) We devise **the first end-to-end trainable neural architecture** EASEE for the active speaker problem (Section 3.1), which learns effective feature representations. (2) In EASEE, we propose **a novel iGNN block** to aggregate spatial and temporal context based on a composition of spatial and temporal message passing. We show this reformulation of the graph structure is key to achieve state-of-the-art results (Section 4.1). (3) Based on EASEE, we propose **the first weakly-supervised ASD approach** that enables the use of only audio labels to generate predictions on visual data (Section 4.3). To ensure reproducible results and foster future research, we have made all the resources of this project available at: `https://github.com/fuankarion/end-to-end-asd`.

2 Related Work

Early approaches to the ASD problem [12] attempted to correlate audiovisual patterns using time-delayed neural networks [42]. Follow up works [37,15] approached the ASD task by limiting the analysis only to visual patterns. These approaches rely only on visual data given the biases of the single speaker scenario (*i.e.* speech can only be attributed to the single visible speaker). A parallel corpus of work focused on the complementary task of voice activity detection (VAD), which aims at finding speech activities among other acoustic events [38,6]. Similar to visual data, audio-only information was also proven to be useful in single speaker scenarios [13].

The recent interest in deep neural architectures [36,27,26] shifted the focus in the ASD problem from hand-crafted feature design to multi-modal representation learning [32]. As a consequence, ASD has become dominated by CNN-based approaches, which rely on convolutional encoders originally devised for image analysis tasks [35]. Recent works [5,11] approached the more general multi-speaker scenario, relying on the fusion of multi-modal information from individual speakers. Concurrent works have also focused on audiovisual feature alignment. This resulted in methods that rely on audio as the primary source of supervision [4], or focused on the design of multi-modal embeddings [10,11,31,39].

The recent availability of large-scale data for the ASD task [35] has enabled the use of state-of-the-art deep convolutional encoders [19,18]. In addition to these deep encoders, current approaches have shifted focus to directly modeling the temporal features over short temporal windows, typically by optimizing a Siamese Network with modality specific streams. The work of Chung *et al.* [9] explored the use of a hybrid 3D-2D encoder pretained on VoxCeleb [10] to analyze these temporal windows, while Zhang *et al.* [49] focused on improving the feature representation by using a contrastive loss [17] between the modalities.

To complement this short-term analysis, many methods[25,28,40] have aimed to incorporate contextual information from overlapping visual tracklets. The work of Alcazar *et al.* [1] introduced a data structure to represent an active speaker scene, and the features in this structure are improved by using self-attention[43,41] and recurrent networks [20].

Current state-of-the-art techniques incorporate contextual representation and rely on deep 3D encoders for the initial feature encoding and recurrent networks or self-attention to analyze the scene's contextual information [28,25,40,50]. We depart from this standard approach and devise a strategy to train end-to-end networks that simultaneously optimize features from a shared multi-modal encoder. This enables the direct optimization of temporal and spatial features for the ASD problem in a multi-speaker setup.

2.1 Graph Convolutional Networks

The current interest in non-Euclidean data [16,21,22,29,30,47] has focused the attention of the research community on Graph Convolutional Networks (GCNs) as an efficient variant of CNNs [24,45]. GCNs have achieved state-of-the-art results in zero-shot recognition [44,23], 3D understanding [16,29,46], and action recognition in video [21,47,48] by harnessing the flexibility of graphs representations. Recently, GCNs have been widely used in the field of action recognition, focusing on skeleton-based approaches that rely only on visual data [2,14]. For applications in audiovisual contexts, GCNs have been utilized to study inter-correlations in videos for automatic recognition of emotions in conversations [33,34]. In the ASD domain, Alcazar *et al.* [28] introduced the use of GCNs, developing a two-stage approach where a GCN network would module interactions between audio and video across multiple frames. We present an alternative to this approach where we focus on the end-to-end modelling, and perform independent steps of message passing along the spatial and temporal dimensions.

3 End-to-End Active Speaker Detection

Our approach relies on the initial generation of independent audio and visual embeddings at specific timestamps. These embeddings are fused and jointly optimized by means of a graph convolutional network [24]. To this end, we devise a neural architecture with three main components: (i) audio Encoder, (ii) visual Encoder, and a (iii) spatio-temporal Module. The visual encoder (f_v) performs multiple forward passes (one for each available tracklet), and the audio encoder (f_a) performs a single forward pass on the shared audio clip. These features are arranged according to their temporal order and (potential) spatial overlap, creating an intermediate feature embedding (Φ) that enables spatio-temporal reasoning. Unlike other methods, we construct Φ such that it can be optimized end-to-end. Thus Φ captures multi-modal and multi-speaker information, enables information flow across modalities, and ultimately improves network predictions. Figure 2 contains an overview of our proposed approach.

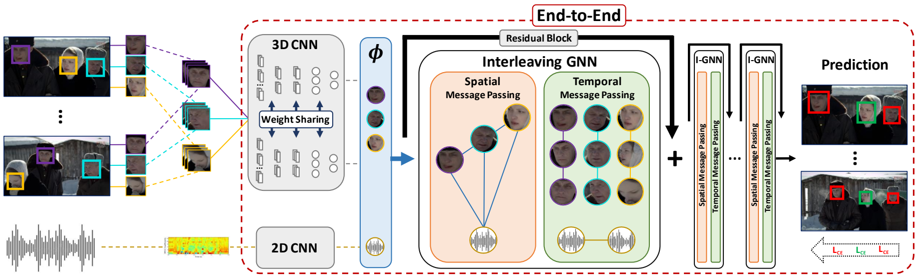

Fig. 2: **Overview of the EASEE architecture.** We fuse information from multiple visual tracklets, and their associated audio track. We rely on a 3D CNN to encode individual face tracklets, and a 2D CNN to encode the audio stream (Grey Encoders). These embeddings are assembled into an initial multi-modal embedding ($\varPhi$) containing audiovisual information from multiple persons in a scene. We map this embedding into a graph structure that performs message passing steps over spatial (light orange) and temporal dimensions (light green). Our layer arrangement favors independent massage passing steps along the temporal and spatial dimensions.

3.1 EASEE Network Architecture

The main goal of EASEE is to aggregate related temporal and spatial information from different modalities over a video segment. To enable efficient end-to-end computation, we do not densely sample all the available tracklets in a temporal window, but rather define a strategy to sub-sample audiovisual segments inside a video. We define a set of temporal endpoints where the original video data (visual and audio) is densely sampled. At every temporal endpoint, we collect visual information from the available face tracklets and sample the associated audio signal (See Figure 3). To further limit the memory usage, we define a fixed number of tracklets (i) to sample at every endpoint. Since the visual stream might contain an arbitrary number of tracklets, we follow [1] at training time and sample i tracklets with replacement. Hence, from every temporal endpoint, we create $i+1$ feature embeddings associated with it (i visual embeddings from f_v and the audio embedding from f_a).

We create temporal endpoints over a video segment following a simple strategy, we select a timestamp t and create l temporal endpoints over the video at a fixed stride of k frames. The location of every endpoint is then given by $L = \{t, t+k, t+2k, ..., t+lk\}$. This reduces the total number of samples from the video data by a factor of k and allows us to sample longer sections of video for training and inference.

Spatio-Temporal Embedding. We build the embedding $\varPhi$ over the endpoint set L. We define the audiovisual embedding e at time t for speaker s as $e_{t,s} = \{f_a(t), f_v(s,t)\}$. Since there may be multiple visible persons at this endpoint (*i.e.* $|s| \geq 1$), we define the embedding for an endpoint at time t with up to i

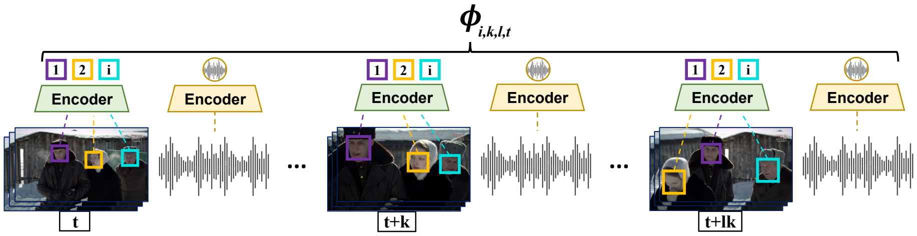

Fig. 3: **EASEE Sub-Sampling.** For every temporal endpoint, we sample i face tracklets and the corresponding audio signal. This sampling is repeated over l consecutive temporal endpoints separated by stride k. The $i+1$ feature embeddings obtained at each timestamp are forwarded through the audio (yellow) and visual (light green) encoders fused into the spatio-temporal embedding $\Phi_{i,k,l,t}$.

speakers as $E_{t,i} = \{e_{t,0}, e_{t,1}, e_{t,2}, ..., e_{t,i}\}$. The full spatio-temporal embedding $\Phi_{i,k,l,t}$ is created by sampling audio and visual features over the endpoint set L, where $\Phi_{i,k,l,t} = \{E_{t,i}, ..., E_{t+k,i}, ..., E_{t+lk,i}\}$. As $\Phi_{i,k,l,t}$ is assembled from independent forward passes of the f_a and f_v encoders, we share weights for forward passes in the same modality, thus each forward/backward pass accumulates gradients over the same weights. This shared weight scheme largely simplifies the complexity of the proposed network, and keeps the total number of parameters stable regardless of the values for l and i.

Upon computing the initial modality embeddings, we map $\Phi_{i,k,l,t}$ into a spatio-temporal graph representation. Following [28], we map each feature in $\Phi_{i,k,l,t}$ into an individual node, resulting in a total of $(i+1)*l$ nodes. Every feature embedding goes through a linear layer for dimensionality reduction before being assigned to a node. Unlike [28], we are not interested in building a unique graph structure that performs message passing over all the possible relationships in the node set. Instead, we choose to independently model the two types of information flow in the graph, namely spatial information and temporal information.

3.2 Graph Neural Network Architecture

In EASEE, the GCN component fuses spatio-temporal information from video segments. This module implements a novel composition pattern where the spatial and temporal information message passing are performed in subsequent layers. We devise a building block (iGNN) where the spatial message passing is performed first, then temporal message passing occurs. After these two forward passes, we fuse the feature representation with the previously estimated feature embedding (residual connection). We define the iGNN block at layer J as:

$$\Phi_s = M^s(A^s\Phi;\theta^s), \Phi_t = M^t(A^t\Phi;\theta^t)$$
$$\Phi^{J+2} = iGNN(\Phi^J) = (M^t \circ M^s)(\Phi^J) + \Phi^J = M^t\big(\underbrace{M^s\,(\Phi^J)}_{\Phi^{J+1}}\big) + \Phi^J$$

Here, M^s is a GCN layer that performs spatial message passing using the spatial adjacency matrix A^s over an initial feature embedding (Φ^J), thus producing an intermediate representation with aggregated local features (Φ^{J+1}). Afterwards the GCN layer M^t performs a temporal message passing using the temporal adjacency matrix A^t. θ^s and θ^t are the parameter set of their respective layers. The final output is complemented with a residual connection, thus favoring gradient propagation.

In EASEE, the assignment of elements from the embedding $\Phi_{i,k,l,s}$ to graph nodes remains stable throughout the entire GCN structure (*i.e.* we do not perform any pooling). This allows us to create a final prediction for every tracklet and audio clip contained in $\Phi_{i,k,l,t}$ by applying a single linear layer. This arrangement creates two types of nodes: *Audio Nodes*, which generate predictions for the audio embeddings (*i.e.* speech detected or silent scene), and *Video Nodes* which generate predictions for the visual tracklets (*i.e.* active speaker or silent). EASEE's final predictions are made only from the output of visual nodes. Audio nodes are supervised in training, but their forward phase output is not suitable for the ASD task. The training loss is defined as: $\mathcal{L} = \mathcal{L}_a + \mathcal{L}_v$. Where $\mathcal{L}_a$ is the loss over all audio nodes and $\mathcal{L}_v$ is the loss over all the video nodes. Both losses are implemented as cross-entropy loss (CE).

3.3 Weakly Supervised Active Speaker Detection

State-of-the-art methods rely on fully supervised approaches to generate consistent predictions in the ASD problem. Typically, they work in a fully supervised manner in both learning stages, using audiovisual labels to train the initial feature encoder and also to supervise the second stage learning [25,40,1,28]. The end-to-end nature of EASEE enables us to approach the active speaker problem from a novel perspective, where the multi-speaker scenario can be analyzed relying on a weak supervision signal, namely audio labels. In comparison to visual labels, audio ground-truth is less expensive to acquire, as it only establishes the start and end point of a speech event. Meanwhile, labels for visual data must establish the fine-grained association between every temporal interval in the speech event and its visual source.

A naive training of EASEE with audio labels only, would optimize the predictions for the audio nodes (speech events). As outlined before, such predictions are suitable for the voice activity detection task, but the more fine grained ASD task will have poor performance as the visual nodes lack any supervision and yield random outputs. To generate meaningful predictions for the visual nodes while relying only on audio supervision, we reformulate our end-to-end training to enforce information flow between modalities by adding two extra loss functions on the graph structure. This reformulation enables meaningful predictions over the visual data despite the lack of visual ground-truth. We name this version of our approach EASEE-W, a novel architecture that is capable of making active speaker predictions that rely only on weak binary supervision labels from the audio stream. An overview of the key differences between EASEE and EASEE-W is show in Figure 4.

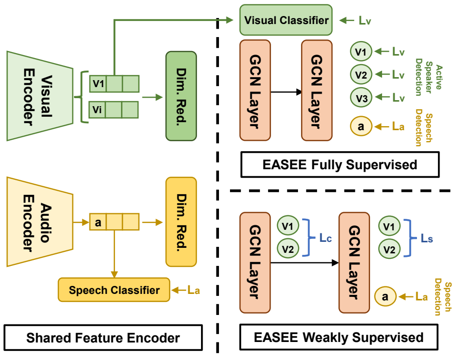

Fig. 4: **EASEE Weakly Supervised.** We drop all the visual supervision ($\mathcal{L}_v$) in EASEE and enforce positive predictions in the video nodes (light green) in the presence of a speech event ($\mathcal{L}_s$), along with consistent visual feature representations for the same identities ($\mathcal{L}_c$).

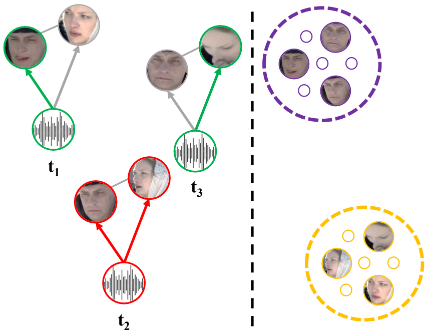

Fig. 5: **Weakly Supervised Losses.** We enforce an individual speaker assignment if there is a detected speech event (left). Temporal consistency pulls together features for faces of the same person and creates differences for faces of different persons (right).

Local assignment loss. We design a loss function that models local dependencies in the ASD problem: if there is a speech event, we must attribute the speech to one of the locally associated video nodes. Let V_t be the output of video nodes at time t ($|V_t| \geq 2$), and y_{at} the ground truth for the audio signal at time t:

$$L_s = y_{at}(y_{at} - \max(V_t)) + (1 - y_{at}) \max(V_t)$$

The first term $y_{at}(y_{at} - \max(V_t))$ will force EASEE-W to generate at least one positive prediction in V_t if $y_{at} = 1$ (*i.e.* select a speaker if speech is detected). Likewise, the second term $(1 - y_{at}) \max(V_t)$ will force EASEE-W to generate only negative predictions in V_t in the absence of speech. While this loss forces the network to generate video labels that are locally consistent with the audio supervision, we show that these predictions only improve the performance over a random baseline and do not represent an improvement over trivial audiovisual assignments.

Visual contrastive loss. We complement L_s with a contrastive loss (L_c) applied over the video data. As shown in Figure 5, the goal of this loss is to enforce feature similarity between video nodes that belong to the same person, and promote feature differences for non-matching identities. Considering that the AVA-ActiveSpeaker dataset [35] does not include identity meta-data, we approximate the sampling of different identities by selecting visual data from concurrent tracklets[3]. To simplify the contrastive learning, we modify the sampling scheme for

[3] Since tracklets include a single face and were manually curated, it is guaranteed that two tracklets that overlap in time belong to different identities. If there is a single

EASEE-W, and force $i = 2$ regardless of the real number of simultaneous tracklets. If there are more than 2 visible persons in the scene, we just sample without replacement.

In practice, we follow [7] and apply this loss on the second to last layer of the iGNN block. Let $\mathcal{L}_a$ be the loss for the audio nodes in the last iGNN block (see Figures 4 and 5), then the loss used for EASEE-W is: $\mathcal{L}_w = \mathcal{L}_a + \mathcal{L}_s + \mathcal{L}_c$. No video labels are required, *i.e.* the speaker-to-speech assignments are unknown.

3.4 Implementation Details

We provide additional training details in the **supplementary material**

4 Experimental Results

In this section, we provide extensive experimental evaluation of our proposed method. We mainly evaluate EASEE on the AVA-ActiveSpeaker dataset [35] and also present additional results on Talkies [28]. We begin with a direct comparison to state-of-the-art methods. Then, we perform an ablation analysis, assessing our main design decisions and their individual contributions to EASEE's final performance. We conclude by presenting the empirical evaluation of EASSE-W in the weakly supervised setup. We include further assessment in known challenging scenarios in the **supplementary material.**

AVA-ActiveSpeaker [35] is the first large-scale test-bed for ASD. AVA-ActiveSpeaker contains 262 Hollywood movies: 120 in the training set, 33 in validation, and the remaining 109 in testing. The dataset provides bounding boxes for a total of 5.3 million faces. These face detections are manually linked to produce face tracks that contain a single identity. All AVA-ActiveSpeaker results reported in this paper were obtained using the official evaluation tool provided by the dataset creators, which uses average precision (mAP) as the main metric for evaluation.

Talkies is a manually labeled dataset for the ASD task [28]. This dataset was collected from social media videos and contains $23,507$ face tracks extracted from a total of 799,446 individual face detections. Unlike AVA-ActiveSpeaker, it is based on short clips and about 20% of the speech events are off-screen speech, *i.e.* the event cannot be attributed to a visible person.

4.1 Comparison to state-of-the-art

We compare EASEE against state-of-the-art ASD methods. The results for EASEE are obtained with $l = 7$ temporal endpoints, $i = 2$ tracklets per endpoint, and a stride of $k = 5$. This configuration allows for a sampling window of about 2.41 seconds regardless of the selected backbone. For fair comparison with

person in the scene, we sample additional visual data from another tracklet in a different movie where no speech event is detected.

	Visual Encoder		Temporal	
Method	**Backbone**	**2D/3D**	**Context**	**mAP**
AVA Baseline *et al.* [35]	MobileNet	2D	✗	79.2
AVA Baseline + GRU *et al.* [35]	MobileNet	2D	✓	82.2
FaVoA [3]	ResNet18	2D	✓	84.7
MAAS-LAN [28]	ResNet18	2D	✗	85.1
Chung et al. [9]	ResNet18	3D+2D	✓	85.5
ASC [1]	Resnet18	2D	✓	87.1
MAAS-TAN [28]	ResNet18	2D	✓	88.8
EASEE-2D (Ours)	ResNet18	2D	✓	91.1
UniCon [50]	Multiple	2D	✓	92.0
Zhang *et al.* [49]	Custom	3D+2D	✗	84.0
EASEE-50 *l=1, i=3* (Ours)	ResNet18	3D	✗	89.6
TalkNet [40]	Custom	3D+2D	✓	92.3
EASEE-18 (Ours)	ResNet18	3D	✓	93.3
ASDNet [25]	ResNext101	3D	✓	93.5
EASEE-50 (Ours)	ResNet50	3D	✓	**94.1**

Table 1: **State-of-the-art Comparison on AVA-ActiveSpeaker.** Our best network (EASEE-50) outperforms any other method by at least 0.6 mAP even approaches that build upon much deeper networks. Our smaller network (EASEE-18) remains competitive with the previous state-of-the-art. In the 2D scenario EASEE-2D only lags behind UniCon [50], improving the closest method by at least 0.9 mAP.

other methods, we report results of three EASEE variants: 'EASEE-50' that uses a 3D backbone based on the ResNet50 architecture, 'EASEE-18' that uses a 3D model based on the much smaller Resnet18 architecture, and 'EASEE-2D' that uses a 2D Resnet18 backbone. Results are summarized in Table 1.

We find that the optimal number of iGNN blocks changes according to the baseline architecture. For the ResNet18 encoder, 6 blocks (24 layers total in the GCN) are required to achieve the best performance, whereas for ResNet50, only 4 blocks (16 layers total in the GCN) are required. Since we find the best results with $i = 2$, and there are scenes with 3 or more simultaneous tracklets, we follow [28]. At inference time, we split the speakers in non-overlapping groups of 2, and perform multiple forward passes until every tracklet has been labeled.

We observe that our method outperforms all the other approaches in the validation subset. EASEE-50 is 0.6 mAP higher than the previous state-of-the-art (ASDNet [25]). We highlight that ASDNet relies on the deep ResNext101 encoder, whereas EASEE-50 is built on the much smaller ResNet50. Our smaller version (EASEE-18) only lags behind ASDNet by 0.2, and outperforms every other model by at least 1.0 mAP. We also implement a version of EASEE-50 that models only spatial relations (*i.e.* $l = 1$). This model reaches 89.6 mAP, outperforming every other network that generates predictions without long-term temporal modelling by at least 4.5 mAP. Finally EASEE-2D outperforms every other 2D approach except UniCon[50], we explain this result as [50] presents a far more complex approach that includes multiple 2D backbones to analyze audiovisual data, scene layout and speaker suppression, along with bi-directional GRUs [8] for temporal aggregation.

4.2 Ablation Study

We ablate our best model (EASEE-50) to assess the individual contributions of our design choices: end-to-end training, iGNN block, and the residual connections between the iGNN blocks. Table 2 contains the individual assessment of each component. The most important architectural design is the end-to-end training, which contributes 1.6 mAP. The proposed iGNN brings about 0.4 mAP when compared against a baseline network where spatial and temporal message passing is performed in the same layer. Finally, residual connections between iGNN blocks contribute with an improved performance of 0.3 mAP.

Intermediate Embedding Configuration. We compare the performance of EASEE-50 with different configurations of the intermediate embedding Φ. In Table 3, we assess the performance of EASEE-50 when changing the number of temporal endpoints l and the number of simultaneous tracklets i. We observe that the best performance arises when $i = 2$, which is in stark contrast to other methods [25,28,50] that often rely on aggregating information from 4 or more visual tracklets. We attribute this to the end-to-end nature of EASEE, where contextual cues are directly optimized for the ASD problem, thus requiring less spatial data for effective predictions. Nonetheless, for small values of l, we find that EASEE actually benefits from a larger number of visual tracklets ($i = 3$). This suggests that in the absence of strong temporal cues, EASEE will focus on extracting meaningful information from the spatially adjacent tracklets.

We also observe that the temporal dimension of the problem (number of endpoints l) is more relevant than the spatial component (number of concurrent tracklets i). When increasing l from 1 to 7, performance improves significantly, by 4.8 mAP on average. In contrast, increasing visual tracklets from $i = 1$ to $i = 4$ only yields 1.1 mAP improvement on average. This is consistent with related works, which show a performance boost when incorporating recurrent Units and long temporal samplings [35,1,25].

			Residual	
Network	**End-to-End**	**iGNN**	**Connections**	**mAP**
EASEE-50	✗	✗	✗	91.9
EASEE-50	✓	✗	✗	93.5
EASEE-50	✓	✗	✓	93.7
EASEE-50	✓	✓	✗	93.8
EASEE-50	✓	✓	✓	**94.1**

Table 2: **AVA-ActiveSpeaker Ablation.** We assess the empirical contribution of the most relevant components in EASEE. Residual connections contribute about 0.3 mAP and the proposed iGNN block 0.4 mAP. Overall the most relevant design choice is the end-to-end trainable nature of EASEE contributing 1.6 mAP.

	End Points (l)				
Speakers (i)	1	3	5	7	9
1	86.6	90.5	92.4	92.9	92.6
2	89.0	92.3	93.4	**94.1**	93.8
3	89.6	92.2	93.3	93.8	93.4
4	89.2	91.8	93.1	93.7	93.2

Table 3: **End Points vs speaker.** Longer temporal windows allow to improve the performance, achieving the best result at $l = 7$. A large number of speakers favors performance in shorter windows but $i = 2$ is the best parameter for long windows ($l \geq 3$).

	End Points (l)				
Clip Size	1	3	5	7	9
11	87.6	91.4	93.1	93.5	93.2
13	88.3	91.7	93.3	93.9	93.6
15	89.0	92.3	93.4	**94.1**	93.8
17	89.3	92.5	93.3	93.9	93.7

Table 4: **End Points vs Input Clip.** Long temporal samplings enables better predictions in most scenarios. In the EASEE architecture the input size for the 3D encoder also provides improved performance, the optimal is 15 frames, which equals to 0.62s

In Table 4, we analyze the effect of the input clip size to the encoder f_v in EASEE. We find that as the clip size increases, performance also improves but saturates around 15 frames (about 0.62 seconds). For every clip size, longer temporal sampling (more endpoints) provides better results. The best result is achieved at $l = 7$ with clips of 15 frames.

Design of iGNN blocks. We assess the effectiveness of the proposed iGNN block by comparing it against the following fusion alternatives: (a) Temporal-Spatial (iGNN-TS), an immediate alternative to iGNN where temporal message passing is performed before any spatial message passing is done; (b) Two Stream, where two independent GCN streams perform spatial and temporal message passing respectively, and these streams are fused at the end of the network; (c) Parallel, where the block performs spatial and temporal message passing in parallel and fuses the features using a fully connected layer; (d) Spatio-Temporal, where a single graph structure performs temporal and spatial message passing at the same time [28]. Table 5 summarizes the results.

Overall, we find that the best block design is the one in which spatial message passing occurs first. Reversing the order of message passing results in a very similar alternative with only minor performance degradation. In comparison, the two-stream approach performs significantly worse than all other alternatives, suggesting that the fusion of temporal and spatial information must occur earlier

iGNN	**iGNN-TS**	**Two Stream**	**Parallel**	**Spatio-temporal** [28]
94.1	94.0	92.8	93.7	93.7

Table 5: **iGNN Layering Strategies.** We compare multiple strategies to assemble our iGNN block, we find that interleaving the temporal and spatial messages brings the best results. In comparison a joint massage passing will reduce the performance by 0.4 mAP, a naive join of this steps with a linear layer reports the same performance reduction.

to be effective in an end-to-end scenario. Joint spatio-temporal messaging also has high performance, but still lags behind the iGNN block.

We conclude this section with the evaluation of EASEE on the Talkies dataset[28]. Here, we test: (i) a direct transfer of EASEE-50 into the validation set of Talkies, (ii) directly training EASEE on Talkies, and (iii) using Talkies as downstream task after pre-training on AVA-ActiveSpeaker. Table 6 summarizes the results. EASEE outperforms [28] for the direct transfer on the Talkies dataset. Moreover, training on Talkies results in a high performance comparable to that of the AVA-ActiveSpeaker dataset, this is particularly interesting as Talkies is a dataset that contains a large portion of scenes with out-of screen speech, a situation that is extremely rare in the AVA-Active Speaker. Finally, the using talkies as a downstream task results in 1.0 mAP improvement, which is about 15% relative error improvement.

4.3 Weak Supervision

We conclude this section by evaluating the weakly supervised version of EASEE, *i.e.* EASEE-W. To the best of our knowledge, there are no comparable methods that strictly rely on weak (audio only) supervision in the ASD task. Therefore, we establish multiple baselines, from random predictions to direct speech-to-speaker assignment.

We first consider baselines that ignore audio labels and the structure of the ASD problem: i) *random* baseline where every speaker gets a random score sampled from a uniform distribution between $[0, 1]$. ii) *Naive Recall* where we trivially predict every tracklet as an active speaker and iii) *Naive Precision* that trivially predicts every tracklet as silent. We also build baselines that rely on audio supervision. We use our trained audio encoder f_a to detect speech intervals and generate random speech-to-speaker assignments within that time window. We explore two approaches: iv) *Naive Audio assignment* where we choose a random visible speaker whenever a speech event is detected. v) *Largest Face Audio assignment* since AVA-Active Speaker is a collection of Hollywood movies, we follow a common bias in commercial movies, and assign the speech event to the tracklet that occupies the largest area in the screen. Table 7 summarizes the results of this experiments in the AVA-ActiveSpeaker dataset.

Network	**AVA Pre-train**	**Talkies Training**	**mAP**
MAAS-TAN [28]	✓	✗	79.1
EASEE-50	✓	✗	86.7
EASEE-50	✗	✓	93.6
EASEE-50	✓	✓	94.5

Table 6: **Evaluation on Talkies Dataset.** We evaluate EASEE on the Talkies dataset. It outperforms the existing baseline on the direct transfer from AVA-ActiveSpeaker, and show the results of training EASEE end-to-end in Talkies. Finally we test the effectiveness of AVA-ActiveSpeakers as pre-training for Talkies.

Network	mAP
Random	25.1
Naive Recall	27.1
Naive Precision	27.1
Naive Audio Assignment	47.7
Large Face Audio Assignment	49.1
EASEE-W L_a only	26.1
EASEE-W L_a, L_c only	25.8
EASEE-W L_a, L_s only	54.4
EASEE-W (L_a, L_s, L_c)	76.2
Fully Supervised 2D encoder [35,1]	79.5

Table 7: **Weak Supervision.** We show that EASEE-W largely improves over baseline approaches for ASD, it outperform a naive baseline by 28.5 mAP, and remains competitive with fully supervised 2D encoder.

We observe that random baselines largely under-perform. Even when the predictions have a bias towards the largest class (silent) results are just 27.1 mAP. A relevant increment in performance (about 20 mAP) appears when the audio supervision is used to generate the naive visual assignments. This improvement is a direct result of the structure in the ASD problem, where speech events are attributed to a defined set of sources.

When we apply EASEE-W, we see the complementary behaviour of the proposed loss functions. The baseline with audio supervision (L_a only) exhibits no meaningful improvement over the random base, despite the GCN structure. A similar situation can be observed if we use the audio supervision and enforce temporal consistency on the visual features (L_c). This indicates that information flow across modalities can not be trivially enforced by the GCN module or temporal visual consistency. Including the assignment loss (L_s) results in a scenario that already improves over the naive assignments suggesting that local attributions already favor the some meaningful audiovisual patterns. Finally, the best result is achieved when assignments and temporal consistency for the visual data are considered. This result improves over any baseline by at least 27 mAP. We conclude this section highlighting that this result is competitive with baseline approaches that rely on encoding short-temporal information from a single speaker as outlined in [35,1].

5 Conclusion

We introduced EASEE, a multi-modal end-to-end trainable network for the ASD task. EASEE outperforms state-of-the-art approaches in the large scale AVA-ActiveSpeaker[35] dataset, and transfers effectively to smaller sets that contain out-of-screen speech. EASEE allows for fully supervised and weakly supervised training by leveraging the inherent structure of the ASD problem and the natural consistency in video data. Future explorations on the ASD problem might rely on our label efficient training setup.

Acknowledgments. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding.

References

1. Alcázar, J.L., Caba, F., Mai, L., Perazzi, F., Lee, J.Y., Arbeláez, P., Ghanem, B.: Active speakers in context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12465–12474 (2020)
2. Cai, J., Jiang, N., Han, X., Jia, K., Lu, J.: Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2735–2744 (2021)
3. Carneiro, H., Weber, C., Wermter, S.: Favoa: Face-voice association favours ambiguous speaker detection. In: International Conference on Artificial Neural Networks. pp. 439–450. Springer (2021)
4. Chakravarty, P., Mirzaei, S., Tuytelaars, T., Van hamme, H.: Who's speaking? audio-supervised classification of active speakers in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 87–90 (2015)
5. Chakravarty, P., Zegers, J., Tuytelaars, T., Van hamme, H.: Active speaker detection with audio-visual co-training. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 312–316 (2016)
6. Chang, J.H., Kim, N.S., Mitra, S.K.: Voice activity detection based on multiple statistical models. IEEE Transactions on Signal Processing **54**(6), 1965–1976 (2006)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
9. Chung, J.S.: Naver at activitynet challenge 2019–task b active speaker detection (ava). arXiv preprint arXiv:1906.10555 (2019)
10. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
11. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision. pp. 251–263. Springer (2016)
12. Cutler, R., Davis, L.: Look who's talking: Speaker detection using video and audio correlation. In: International Conference on Multimedia and Expo (2000)
13. Ding, S., Wang, Q., Chang, S.y., Wan, L., Moreno, I.L.: Personal vad: Speaker-conditioned voice activity detection. arXiv preprint arXiv:1908.04284 (2019)
14. Duhme, M., Memmesheimer, R., Paulus, D.: Fusion-gcn: Multimodal action recognition using graph convolutional networks. arXiv preprint arXiv:2109.12946 (2021)
15. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automated naming of characters in tv video. Image and Vision Computing **27**(5), 545–559 (2009)
16. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. arXiv preprint arXiv:1906.02739 (2019)
17. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
18. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
21. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5308–5317 (2016)
22. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1219–1228 (2018)
23. Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., Xing, E.P.: Rethinking knowledge graph propagation for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11487–11496 (2019)
24. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
25. Köpüklü, O., Taseska, M., Rigoll, G.: How to design a three-stage architecture for audio-visual active speaker detection in the wild. arXiv preprint arXiv:2106.03932 (2021)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
27. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems **2** (1989)
28. León-Alcázar, J., Heilbron, F.C., Thabet, A., Ghanem, B.: Maas: Multi-modal assignation for active speaker detection. arXiv preprint arXiv:2101.03682 (2021)
29. Li, G., Qian, G., Delgadillo, I.C., Müller, M., Thabet, A., Ghanem, B.: Sgas: Sequential greedy architecture search (2019)
30. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 335–351 (2018)
31. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
32. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
33. Nie, W., Ren, M., Nie, J., Zhao, S.: C-gcn: Correlation based graph convolutional network for audio-video emotion recognition. IEEE Transactions on Multimedia (2020)
34. Ren, M., Huang, X., Li, W., Song, D., Nie, W.: Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. IEEE Transactions on Multimedia (2021)
35. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., 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). pp. 4492–4496. IEEE (2020)
36. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
37. Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., Darrell, T.: Visual speech recognition with loosely synchronized feature streams. In: ICCV (2005)
38. Tanyer, S.G., Ozer, H.: Voice activity detection in nonstationary noise. IEEE Transactions on speech and audio processing **8**(4), 478–482 (2000)

39. Tao, F., Busso, C.: Bimodal recurrent neural network for audiovisual voice activity detection. In: INTERSPEECH. pp. 1938–1942 (2017)
40. Tao, R., Pan, Z., Das, R.K., Qian, X., Shou, M.Z., Li, H.: 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 (2021)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE transactions on acoustics, speech, and signal processing **37**(3), 328–339 (1989)
43. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
44. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6857–6866 (2018)
45. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., Weinberger, K.: Simplifying graph convolutional networks. In: International conference on machine learning. pp. 6861–6871. PMLR (2019)
46. Xie, Z., Chen, J., Peng, B.: Point clouds learning with attention-based graph convolution networks. arXiv preprint arXiv:1905.13445 (2019)
47. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10156–10165 (2020)
48. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
49. Zhang, Y.H., Xiao, J., Yang, S., Shan, S.: Multi-task learning for audio-visual active speaker detection
50. Zhang, Y., Liang, S., Yang, S., Liu, X., Wu, Z., Shan, S., Chen, X.: Unicon: Unified context network for robust active speaker detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3964–3972 (2021)