The Right to Talk: An Audio-Visual Transformer Approach

Thanh-Dat Truong[1,*], Chi Nhan Duong[2,*], The De Vu[3], Hoang Anh Pham[3]
Bhiksha Raj[4], Ngan Le[1], Khoa Luu[1]
[1]CVIU Lab, University of Arkansas [2]Concordia University
[3]VinAI Research [4]Carnegie Mellon University
{tt032, thile, khoaluu}@uark.edu, dcnhan@ieee.org, {v.devt, v.anhph18}@vinai.io, bhiksha@cs.cmu.edu

Abstract

Turn-taking has played an essential role in structuring the regulation of a conversation. The task of identifying the main speaker (who is properly taking his/her turn of speaking) and the interrupters (who are interrupting or reacting to the main speaker's utterances) remains a challenging task. Although some prior methods have partially addressed this task, there still remain some limitations. Firstly, a direct association of Audio and Visual features may limit the correlations to be extracted due to different modalities. Secondly, the relationship across temporal segments helping to maintain the consistency of localization, separation and conversation contexts is not effectively exploited. Finally, the interactions between speakers that usually contain the tracking and anticipatory decisions about transition to a new speaker is usually ignored. Therefore, this work introduces a new Audio-Visual Transformer approach to the problem of localization and highlighting the main speaker in both audio and visual channels of a multi-speaker conversation video in the wild. The proposed method exploits different types of correlations presented in both visual and audio signals. The temporal audio-visual relationships across spatial-temporal space are anticipated and optimized via the self-attention mechanism in a Transformer structure. Moreover, a newly collected dataset is introduced for the main speaker detection. To the best of our knowledge, it is one of the first studies that is able to automatically localize and highlight the main speaker in both visual and audio channels in multi-speaker conversation videos.

1. Introduction

Although human beings possess capabilities of localizing and separating sounds from noisy environments, we still have trouble following a conversation with noises, background voices, or interruptions from other speakers. Either with blind audio separation [42, 53, 62, 60, 70] or visual-aid audio separation [4, 6, 13, 31, 33, 35, 39, 48, 43, 57, 61, 65, 66] approaches, this outlier separation task still remains a challenge in the wide conditions beyond the lab settings. The problem becomes especially harder when dealing with unknown numbers of speakers in an audio. Nachmani et al. [54] make a comparison between methods and show how hard it is to separate voices when the number of sound sources increases. Existing methods achieve high performance with inputs from multiple microphones. Some methods assume a clean set of single source audio examples are available for supervision [2, 28, 71, 72]. In practice, rather than solely trying to separate voices of all speakers in a conversation and determining "*who-spoke-when*", we tend to give more attentions to the ***main speaker***, i.e. *who is on his/her turn of speaking and his/her talk is the main channel of communication*, and ignore the voices of remaining speakers, i.e. ***interrupters*** or ***listener***, or background noises.

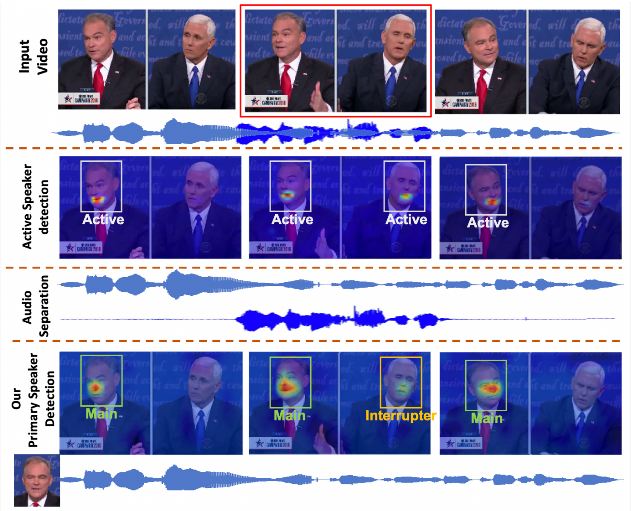


Figure 1. Given a multi-speaker video, our Audio-Visual Transformer can localize and highlight the main speaker in both visual and audio channels. (**Best viewed in color**)

[*] denotes equal contributions

Table 1. Comparisons of our proposed approach and other modeling methods. Sound Source Localization (SSL)

	Ours	LWTNet[5]	SyncNet [14]	SoundOfPixel [72]	CocktailParty [28]
Goal	**Main Speaker Highlight**	Active Speaker Highlight	SSL	Audio Separation	Audio Separation
Temporal Model	**Across-Segments**	Within-Segment	Within-Segment	Within-Segment	Within-Segment
People-Independent	✓	✓	✗	✗	✓
Visual Context modeling (Visual-visual attention)	✓	✗	✗	✗	✗
Audio-Visual Correlation	**Audio-Visual Transformer**	Cosine Distance	Audio-Visual Synchronization	Feature Concatenation	Feature Concatenation

Thus, an approach that highlights the main speaker in visual and audio channels would give new opportunities to popular applications such as auto-muting in a tele-conference or main speaker refocusable video generation.

Given a video of multi-speaker conversation, our goal is to learn an audio-visual model that enables the capabilities of both (1) localizing the main speaker; (2) true cancellation of audio sources of interrupters or background noises; and (3) automatically switching to a new subject when the speakers change their roles. The interruptions from other subjects and the background are considered as noises and removed. In the scope of this work, we focus on ***turn-taking conversation*** as the turn-taking mechanism has been commonly adopted for structuring conversation in social interactions. A subject is considered as the main speaker when he/she properly takes the turn of solo speaking and will continue the talk even after a simultaneous speech occurs [16].

Previous approaches have partially addressed this problem and can be divided into two categories, i.e. *audio-visual synchronization* [7, 14, 15, 56, 44] and *mix-and-separate* [2, 18, 28, 32, 34, 40, 46, 71, 72, 73]. The former exploits the synchronization between audio and video frames within a specific time window to localize the image regions that are more sensitive to audio changes. Meanwhile, the latter learns to separate the speakers' voices from a mix utterance based on audio and visual features. In both cases, there remain some limitations. Firstly, audio-visual relationships are extracted via a concatenation operator or the cosine distance metric. However, as audio and visual features distribute in two different latent spaces by their nature, these methods may not maximize correlations between the two feature domains. Secondly, audio-visual relationships are only considered within a video segment, i.e. a short time window, while ignoring the ones across temporal segments, which helps to maintain the consistency of localization and separation, and *contextual* main-subject switching. Finally, the interactions between subjects in the temporal dimension to deliver accurate tracking and anticipatory decisions about transition to a new target are still ignored.

Contributions. This work introduces a novel Audio-Visual Transformer approach, a cross-modality temporal-based computer vision algorithm, to highlight main speaker in both audio and visual channels (Fig. 1). The contributions of this work are four-fold. (1) The proposed approach exploits various correlations presented in visual and audio signals including "virtual" interactions between speakers in a video scene and relationships between visual and auditory modalities. (2) Rather than extracting audio-visual correlations within a video segment, relationships across segments are further exploited via a temporal self-attention mechanism in the proposed Transformer structure. This helps to engage the contextual information and enhance attentions with longer context so that the main speaker can be robustly identified. (3) A Cycle Synchronization Loss is introduced to learn the main speaker localization in a self-supervised manner. (4) A newly dataset[*] is collected for the main speaker detection. To the best of our knowledge, it is one of the first works that is able to automatically localize and highlight the main speaker in multi-speaker conversation videos on both visual and audio channels (Table 1).

2. Related Work

Active Speaker Localization. This problem aims to localize the sources of sounds in a given video. Some early methods [9, 37, 41, 45, 65] localize the sources of human voice in a video using statistical models and audio-visual correlations. Fisher et al. [30] introduces a multi-media fusion method in a complex domain to capture latent audio-visual relationships. Later, deep learning approaches [14, 56] come into place and exploit the synchronization between visual and audio signals to find the regions in the images that are sensitive to the audio features. Afouras et al. [5] propose LWTNet that extends the synchronize cues with optical flow technique to extract and track audio-visual objects for the localization process. Unlike previous approaches, our work goes one step further by taking into account the *context* of audio-visual features that presents across video segments. Our method does not naively localize all the areas containing voice, but it is able to notice the sound and highlight the location of the *main speaker* in a conversation.

Speaker Audio Separation. Prior methods [42, 53, 62] use audio-only features, i.e characteristics of voice, to resolve this problem. Hershey et al. [38] formulate this task as a clustering problem where the objective is to learn an embedding for each time-frequency element in the spectrogram, such that each embedding cluster associates with a

[*] The dataset and implementation will be publicly available at `https://github.com/uark-cviu/Right2Talk`

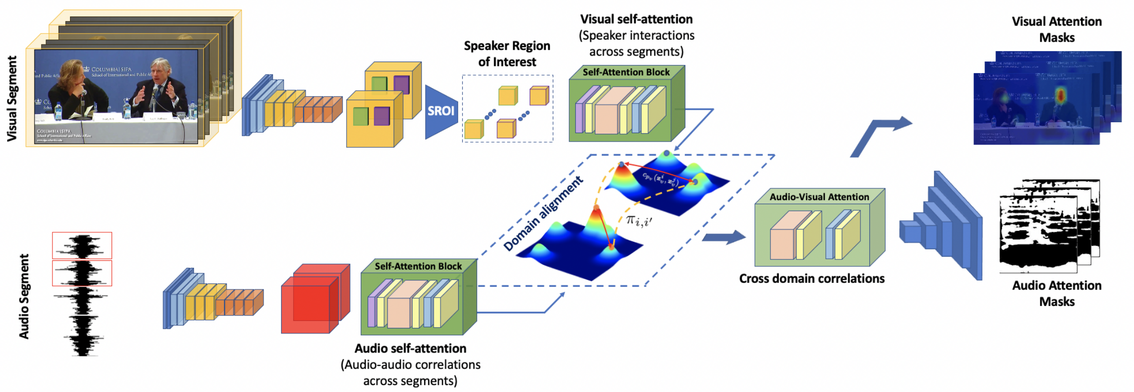


Figure 2. **Our Proposed Audio-Visual Transformer Framework.** Given a video segment set, the context in the conversation is captured with three types of correlations, i.e. visual-visual, audio-audio, and audio-visual attentions. Then, the audio-visual attentional features are adopted for the main speaker localization and audio separation. (**Best viewed in color with 2x zoom in**)

voice of a subject. Zhao et al. [71, 72] detect objects in one or multiple frames and use their appearance and motion to differentiate sounds of objects. Gao et al. [34] propose a co-separation training objective to learn audio-source separation from unlabeled videos containing multiple sources of sounds. Ephrat et al. [28] contribute a large-scale dataset, namely AVspeech, and propose an end-to-end audio-visual architecture. Afouras et al. [2] propose to use the lip regions and consider both audio magnitudes and phases. The aforementioned methods ignore the *context* of the video which is a very important cue for the network to improve the quality of separated voices. Our novel architecture is proposed to to gain the *context* information .

3. The Proposed Method

This work focuses on turn-taking conversations composing talking turns. The length of each turn is flexible according to the conversation's context and contents. Let $\mathbf{x} \in \mathcal{X}$ be a multi-speaker conversation video consisting of a *visual component* $\mathcal{V}$ (a sequence of RGB frames) and an *audio component* $\mathcal{A}$ (a mixed audio of one or multiple speakers).

3.1. The Turn-taking Conversation

With turn-taking regulations, a conversation $\mathbf{x}$ can be decomposed into speaking turns, i.e. $\mathbf{Turn}^h, h = 1, ..., H$ where H is the number of turns in $\mathbf{x}$. Although many speakers can have their voices overlapped during a speaking turn in either cooperative or competitive manner, the role of each speaker can be classified into two groups.

Main Speaker S_m^h**.** A subject S is the main speaker of $\mathbf{Turn}^h$ when he/she carries the conversation and drives it forward [16]. Even an interruption (i.e. simultaneous speech) occurs, the subject will continue to solo speak until the end of the turn. Thus, a long solo speaking of S during $\mathbf{Turn}^h$ can provide an indication for the main speaker role.

Interrupter or Listener S_I^h**.** An interrupter or Listener is the one who has reactions or comments to the main speaker's utterances. These reactions usually occur in a short time window during $\mathbf{Turn}^h$ and end up with the continuation of the main speaker's talk. When the interrupter continues to solo speak after a simultaneous speech, a turn changing of the main speaker occurs. Fig. 3 illustrates an example of speakers' roles in a turn-taking conversation.

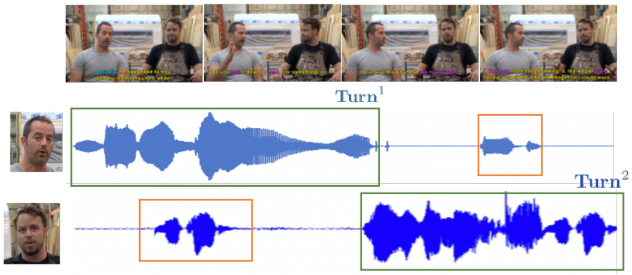


Figure 3. **Turn-taking Conversation.** The conversation is decomposed in to turns where each speaker acts as the main speaker (green box) of a turn, and the other speakers are considered as interrupter or listener (orange box) during that turn.

3.2. Problem Definition

Rather than decomposing $\mathbf{x}$ into $\{\mathbf{Turn}^h\}_1^H$, we propose to present $\mathbf{x}$ as a composition of K segments $\mathbf{Seg}^k = \{\mathbf{v}^k, \mathbf{a}^k\}, k = 1, ..., K$, $\mathbf{v}^k \in \mathcal{V}$ and $\mathbf{a}^k \in \mathcal{A}$. The h-th speaking turn $\mathbf{Turn}^h$ consists of one or multiple segments, i.e. $\mathbf{Turn}^h = \{\mathbf{Seg}^k\}_{h_{start}}^{h_{end}}$ where h_{start} and h_{end} mark the indices of the starting and ending time of h-th turn, respectively. Let S_m^k be the main speaker and S_I^k be the interrupters of $\mathbf{Seg}^k$ in the conversation. We have $S_m^k \equiv S_m^h$ and $S_I^k \equiv S_I^h$ when $\mathbf{Seg}^k \in \mathbf{Turn}^h$. Then, the goal is to extract the location and clean voice of S_m^k for each $\mathbf{Seg}^k$ in the conversation. Formally, the objectives are to learn the visual location map $\mathbf{M}_v^k$ and audio mask $\mathbf{M}_a^k$ of S_m^k as.

$$\begin{aligned} \mathbf{M}_v^{k*} = \arg\min_{\mathbf{M}_v^k} \Big[&-\log P\left(\mathbf{M}_v^k[Loc(S_m^k, \mathbf{v}^k)] | \mathbf{Seg}^{1:k}\right) \\ &+ \log P\left(\mathbf{M}_v^k[Loc(S_I^k, \mathbf{v}^k)] | \mathbf{Seg}^{1:k}\right) \Big] \end{aligned} \tag{1}$$

$$\mathbf{M}_a^{k*} = \arg\min_{\mathbf{M}_a^k} \|\mathbf{M}_a^k \odot Spec(\mathbf{a}^k) - Spec(\mathbf{a}_{S_m^k})\|_1 \tag{2}$$

where $Loc(S_m^k, \mathbf{v}^k)$ is the location of S_m^k in $\mathbf{v}^k$; $Spec(\cdot)$ is the spectrogram conversion operator; $\odot$ is the Hadamard product; and $\mathbf{a}_{S_m^k}$ is the clean voice of S_m^k. $\mathbf{Seg}^{1:k}$ denotes the temporal information provided from the beginning to the k-th segment of the video $\mathbf{x}$. The conditional term indicates the temporal constraint being considered.

To effectively estimate $\mathbf{M}_v^k$ and $\mathbf{M}_a^k$, we propose an Audio-Visual Transformer approach (see Fig. 2) consisting of three learning stages: (1) Learning the context with visual and audio self-attention; (2) Audio-Visual Correlation Learning; and (3) Main speaker localization and audio separation with Conversation Grammar. The proposed Audio-Visual Transformer is formulated via $\{D, \phi, E_v, E_a\}$ as:

$$\begin{aligned} G &= [D \circ \phi](\mathbf{z}_v, \mathbf{z}_a) \\ \mathbf{z}_v^k &= E_v(\mathbf{v}^k|\mathbf{Seg}^{1..k}) \\ \mathbf{z}_a^k &= E_a(\mathbf{a}^k|\mathbf{Seg}^{1..k}) \end{aligned} \tag{3}$$

where E_v and E_a map $\mathbf{v}^k$ and $\mathbf{a}^k$ to their latent representations; and $\circ$ is the functional composition. ϕ is the projection function to the shared representation space where these modalities are comparable. D maps these deep representations to the audio-visual mask of the main speaker.

3.3. Visual and Audio Contextual Learning

Besides prior works on temporal learning [19, 20, 22, 25, 26, 27, 69], in this work, given a conversation, contextual information can be extracted from visual and audio signals, i.e. *visual-visual* and *audio-audio* correlations *across video segments*. While the former assists to track behaviors and interactions of each speaker over the spatial-temporal dimension, the latter provides more cues about the conversation flow, i.e. when and how the main speaker switch his/her role. For example, the higher audio-audio correlation between two or more segments is, the lower the possibility is of the main speaker being switched. Thus, these cross-segment correlations can implicitly embed turn changing signals of the main speaker, and help to avoid the stage of pre-decomposing $\mathbf{x}$ into speaking turns $\mathbf{Turn}^h$. Even when interrupters dominate the main speaker's voice in a certain segment, this "temporal-based" correlation can exploit the relations to previous segments and identify the main speaker. We model the contextual correlations via two encoder structures with a self-attention mechanism before embedding their cross-domain correlations.

3.3.1 Visual-Visual Self-Attention.

Given a sequence of segments $\{\mathbf{v}^k\}_1^K$, the visual encoder E_v consists of three main functions, i.e. feature embedding, self-attention, and feature refinement with attention. Particularly, each $\mathbf{v}^k$ is firstly embedded into a deep feature embedding via $F_v : \mathcal{V} \mapsto \mathcal{F}_v$ as $\mathbf{f}_v^k = F_v(\mathbf{v}^k)$.

Speaker Region of Interest (SROI). Rather than embedding each $\mathbf{v}^k$ into a single feature for attention computation, we propose to project $\mathbf{f}_v^k$ into regions of interest where each region represents a speaker's location in a visual segment and learn the correlations among them. Particularly, let $\mathbf{b} = \{\mathbf{b}_i^k\}, i = 1..N, k = 1..K$ where $\mathbf{b}_i^k \in \mathcal{B} \subset \mathbb{R}^4$ denotes the location of i-th speaker in $\mathbf{Seg}^k$, and N is the number of speakers. The projection function $R : \mathcal{F}_v \times \mathcal{B} \mapsto \mathcal{F}_v$ is defined as $\mathbf{f}_v^{k,i} = R(\mathbf{b}_i^k, \mathbf{f}_v^k)$. We adopt ROI Align [36] for the function R. There are two approaches to obtain $\mathbf{b}$ and N, i.e. face detection and block decomposition. The former adopts a face detection to extract faces in all segments. The latter uniformly decomposes a visual segment into $N = n \times n$ blocks for $\mathbf{b}$. While face detection approach tends to give more direct focus on face regions, our experiments show that block decomposition can provide attentions to regions of face track of the same speaker across segments.

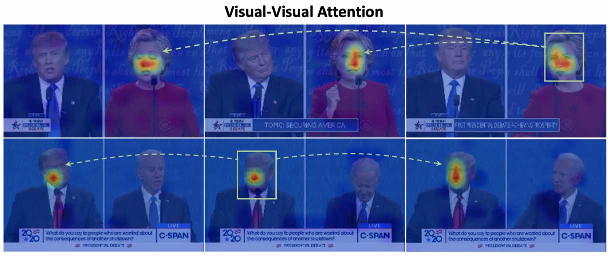


Figure 4. **Visual-Visual Attention.** The attention masks across video segments corresponding to the speaker in the green box. This type of attention can help to track the behaviors and interactions of each speaker over the spatial-temporal dimension.

Virtual Interaction Attention. Given a feature set $\mathbf{f}_v^{k,i}$, the visual-visual context across the spatial-temporal dimension can be expressed as building a dynamic dictionary per feature set with three basic attention based elements [55, 59], i.e. *key, query, value*. While *key* and *query* are trained to support the dictionary look-up process where query feature is highly correlated to its matching key and dissimilar to others, *value* represents a discriminative feature for each speaker. Particularly, the self-attention set $\{\mathbf{k}_v^{k,i}, \mathbf{q}_v^{k,i}, \mathbf{v}_v^{k,i}\}$ is extracted via three learnable projections $\{\Omega_v^Q, \Omega_v^K, \Omega_v^V\}$ as $\mathbf{q}_v^{k,i} = \Omega_v^Q(\mathbf{f}_v^{k,i}); \mathbf{k}_v^{k,i} = \Omega_v^K(\mathbf{f}_v^{k,i}); \mathbf{v}_v^{k,i} = \Omega_v^V(\mathbf{f}_v^{k,i})$. The visual correlation among speakers can be defined as.

$$\alpha_v^{ki,k'j} = \sigma\left(\mathbf{q}_v^{k,i}(\mathbf{k}_v^{k',j})^\top/\sqrt{d}\right) \tag{4}$$

where d is the feature dimension, k' is a segment indexing variable. We consider the attention as a probability distribution that illustrates the responsive attention among speakers. Therefore, the softmax function can be adopted for $\sigma(\cdot)$.

Feature Refinement with attention. With these correlations, the visual self-attention among speakers allows every speaker correlates to all other speakers through the spatial-time dimension. Then, the virtual interaction over speakers is explicitly embedded to their representations as.

$$\mathbf{z}_v^{k,i} = \eta_v\left(\mathbf{f}_v^{k,i} + \sum_{k'=1}^{K}\sum_{j=1}^{N}\alpha_v^{ki,k'j}\mathbf{v}_v^{k,i}\right) \tag{5}$$

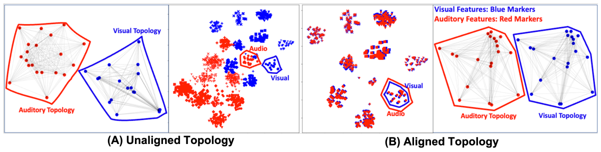


Figure 5. Topology of audio and visual feature domains.

where η_v is the a residual-style MLP. Throughout this process, the features of each speaker in one visual segment can interactively embed in their latent representation the correlations with those of the same speaker in other segments as well as other speakers of the same segment. Fig. 4 illustrates the attention mask across video segments corresponding to the speaker in the green box.

3.3.2 Audio-Audio Self-Attention

Similar to the virtual interaction, the audio self-attention is modeled as the correlation among audio segments. Particularly, let $F_a : \mathcal{A} \rightarrow \mathcal{F}_a$ be an audio embedding function that extracts audio latent representation for audio segments $\{\mathbf{a}^k\}_1^K$ and $\mathbf{f}_a^k = F_a(\mathbf{a}^k)$. The audio self-attention correlation among segments can be computed as in Eqn. (6).

$$\begin{aligned}\mathbf{q}_a^k = \Omega_a^Q(\mathbf{f}_a^k); \mathbf{k}_a^k &= \Omega_a^K(\mathbf{f}_a^k); \mathbf{v}_a^k = \Omega_a^V(\mathbf{f}_a^k)\\ \alpha_a^{k,k'} &= \sigma\left(\mathbf{q}_a^k(\mathbf{k}_a^{k'})^\top/\sqrt{d}\right)\\ \mathbf{z}_a^k &= \eta_a(\mathbf{f}_a^k + \sum_{k'=1}^{K} \alpha_a^{k,k'} \mathbf{v}_a^{k'})\end{aligned} \tag{6}$$

The extracted audio feature of each segment is able to embed the correlation with other audio segments through time.

3.4. Audio-Visual Correlation Learning

The audio-visual correlations are computed from features of two different domains. As ***topologies***, i.e. *how features distributed in the latent space and the correlations among its features* (see Fig. 5), of these modalities may differ significantly, directly associating these features for correlation learning is not efficient. One solution is to set up two encoders E_v and E_a extracting features from latent spaces of the same dimension to leverage the domain differences. However, the topology difference between these modalities may still present. To mitigate this issue, we align two domains' topologies before learning the correlations between their features. By this way, $\mathbf{z}_a^k$ and $\mathbf{z}_v^k$ are well aligned and their correlations can be fully exploited.

Cross Domain Alignment as Optimal Transport (OT) Problem. We present the distributions of visual and audio features by two distributions p_v and p_a where $\mathbf{z}_v \sim p_v(\mathbf{z}_v)$, and $\mathbf{z}_a \sim p_a(\mathbf{a}_a)$; and propose a two-stage alignment process: (1) Sample association between visual and audio samples via transport function π and (2) Topology synchronization. Formally, let π be the transport function where $\pi_{i,i'} = \pi(\mathbf{z}_v^i, \mathbf{z}_a^{i'})$ indicates the probability of association between a visual sample $\mathbf{z}_v^i$ and an audio sample $\mathbf{z}_a^{i'}$. In addition, let $c_{p_v}(\cdot,\cdot)$ and $c_{p_a}(\cdot,\cdot)$ be the cost functions defined as the distance between two samples in visual and audio spaces, respectively. The alignment process is formulated with Gromov-Wasserstein distance as shown in Eqn. (7).

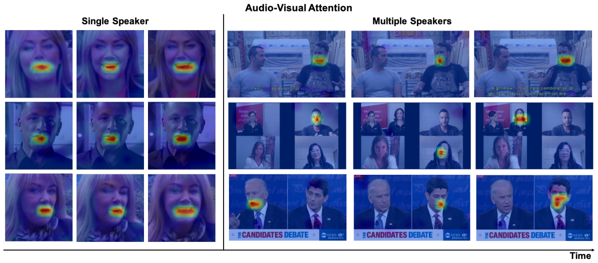


Figure 6. **Audio-Visual Attention To The Main Speaker.** The audio-visual attention mask across segments illustrates the response of audio to the visual. (**Best viewed in color**)

$$\begin{aligned}\mathcal{L}_{align} &= GW^2(c_{p_v}, c_{p_a}, p_v, p_a) = \min_{\pi \in \Pi(p_v, p_a)} J(c_{p_v}, c_{p_a}, \pi)\\ J(c_{p_v}, c_{p_a}, \pi) &= \sum_{i,j,i',j'} |c_{p_v}(\mathbf{z}_v^i, \mathbf{z}_v^j) - c_{p_a}(\mathbf{z}_a^{i'}, \mathbf{z}_a^{j'})|^2 \pi_{i,i'}\pi_{j,j'}\end{aligned} \tag{7}$$

Intuitively, minimizing $J(c_{p_v}, c_{p_a}, \pi)$ aims at finding an appropriated association (i.e. via π) between samples in the two domains as well as minimizing the topology difference between them (i.e. via c_{p_v}, c_{p_a}). Notice that directly solving Eqn. (7) is costly due to the non-convex Quadratic Problem with the time complexity is $O(n^3)$. Therefore, we adopt the the sliced approach [67] for a fast computation of $\mathcal{L}_{align}$. Fig. 5 illustrates visual (blue points) and auditory features (red points) extracted from 500 clip segments of 10 different speakers (e.g. denoted by different markers) and projected into the 2D space using t-SNE method. Thanks to $\mathcal{L}_{align}$ in the alignment stage, visual and auditory features are brought into similar distributions (Fig. 5 (B) (left)) with more aligned feature distributions (Fig. 5 (B) (right)).

Audio-Visual Correlation. With the aligned visual and audio features, we further adopt similar attention mechanism to learn the associations between the visual features of each SROI $\mathbf{z}_v^{k,i}$ and the audio features $\mathbf{z}_a^k$ in each segment as.

$$\begin{aligned}\mathbf{q}^k = \Omega^Q(\mathbf{z}_a^k); \mathbf{k}^{k,i} &= \Omega^K(\mathbf{z}_v^{k,i}); \mathbf{v}^{k,i} = \Omega^V(\mathbf{z}_v^{k,i})\\ \alpha^{k,k'i} &= \sigma\left(\mathbf{q}^k(\mathbf{k}^{k',i})^\top/\sqrt{d}\right)\\ \mathbf{z}^k = \phi(\mathbf{z}_a^k, \mathbf{z}_v^k) &= \eta\left(\mathbf{z}_a^k + \sum_{k'=1}^{K}\sum_{i=1}^{N} \alpha^{k,k'i}\mathbf{v}^{k',i}\right)\end{aligned} \tag{8}$$

The attention matrix assesses how much an audio responds to an SROI in the spatial-temporal dimension. A high response indicates a high correlation between the audio and the speaker associated with that SROI. This association embeds the probability of a speaker to be an active speaker of the audio segment. Fig. 6 illustrates audio-visual attentions in both single and multiple speaker conversation.

3.5. Main Speaker Localization and Audio Separation with Conversation Grammar

Given the audio-visual attentional features $\mathbf{z}^k$ from previous step, the audio-visual masks are computed as follows,

$$
\begin{aligned}
\mathbf{M}_v^k(x,y) &= \begin{cases} \alpha^{k,ki} & \text{if } (x,y) \in \mathbf{b}_i^k, i = 1..N \\ 0 & \text{otherwise} \end{cases} \\
\mathbf{M}_a^k &= D(\mathbf{z}^k)
\end{aligned} \tag{9}
$$

where D is a learnable decoder that maps $\mathbf{z}^k$ to the target audio mask; and $\alpha^{k,ki}$ denotes the correlation score between the audio and SROI of i-th speaker in the k-th segment. The objective functions of Eqns. (1) and (2) to learn $\mathbf{M}_a^k$ and $\mathbf{M}_v$ can be reformulated as follows,

$$
\begin{aligned}
\mathcal{L}_{visual} &= \mathbb{E}\left[-\log \frac{e^{\alpha^{k,ki}}}{e^{\alpha^{k,ki}} + \sum_{j \neq i} e^{\alpha^{k,kj}}} \right] \\
\mathcal{L}_{audio} &= \mathbb{E}\left[||\mathbf{M}_a^k \odot Spec(\mathbf{a}^k) - Spec(\mathbf{a}_{S_m^k})||_1 \right]
\end{aligned} \tag{10}
$$

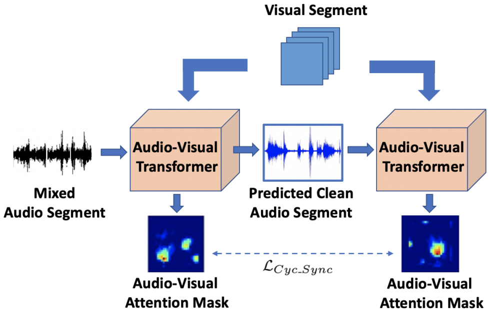
Intuitively, on one hand, $\mathcal{L}_{audio}$ optimizes the model toward voice of the target (Main) speaker. On the other hand, $\mathcal{L}_{visual}$ aims at increasing the correlations between the audio and SROI of the target speaker, while reducing the correlation with other SROIs in spatial-temporal dimensions.

Self-supervised Learning. While our goal is to develop a self-supervised model that learns to localize the main speaker, the ground truth location for Main speaker is absent during the training stage. Therefore, we further propose a self-supervised version of $\mathcal{L}_{visual}$, namely *Cycle Synchronization Loss*, defined as follows,

$$
\begin{aligned}
\mathcal{L}_{Cyc\_Sync} &= \mathbb{E}\left[||\alpha^{k,ki} - \hat{\alpha}_{\max}^{k,ki}||_1 \right] \\
\hat{\alpha}_{\max}^{k,ki} &= \begin{cases} \hat{\alpha}^{k,ki} & \text{if } (x,y) \in \mathbf{b}_{i^*}^k \\ 0 & \text{otherwise} \end{cases} ; i^* = \arg\max_i \hat{\alpha}^{k,ki}
\end{aligned} \tag{11}
$$

where $\hat{\alpha}^{k,ki}$ is the correlation between the predicted clean voice of the target speaker. The intuition of $\mathcal{L}_{Cyc\_Sync}$ is illustrated in Fig. 7 where the goal is to penalize the consistency between two terms: (1) the correlations of the input (mix) voices $\mathbf{a}^k$ and the visual component; and (2) the maximum correlations of the predicted (clean) voice of the target and the visual component. As clean voice is of a single speaker and it reflects similar linguistic content as visual features of the target speaker, its correlations with the visual component can efficiently act as the guidance for localization process. Moreover, by considering only the maximum $\hat{\alpha}^{k,ki}$ in $\hat{\alpha}_{\max}^{k,ki}$, the correlation between the visual of other speakers (i.e. interrupter) and audio is also minimized.

Learning with Conversation Grammar. We adopt the mix-and-separate strategy [28, 72] to obtain the ground truth for audio separation task of $\mathcal{L}_{audio}$ and further extend it with two types of Conversation Grammar, i.e. cooperative and competitive modes. In the first type, each speaker takes turn to speak during the conversation and the role changing happens when a speaker finishes his/her speech. In the second type, mixing voices happen during the interruption of other speakers. From these grammars, we synthesize a video training set containing multiple speakers by (1) randomly selecting different videos in the single subject training set; (2) concatenating these videos sequentially (i.e. cooperative mode); (3) mixing their voices in a short time window and vertically concatenating the video frames (i.e. competitive mode). In all cases, S_m^k is set to the one who occupies the audio segment or the all segments of the whole video, accordingly. The Audio-Visual Transformer is optimized as:



Visual Segment
Audio-Visual Transformer
Mixed Audio Segment
Predicted Clean Audio Segment
Audio-Visual Transformer
$\mathcal{L}_{Cyc\_Sync}$
Audio-Visual Attention Mask
Audio-Visual Attention Mask

Figure 7. **Cycle Synchronization Loss.**

$$
\mathcal{L} = \alpha_{align}\mathcal{L}_{align} + \alpha_{visual}\mathcal{L}_{Cyc\_Sync} + \alpha_{audio}\mathcal{L}_{audio} \tag{12}
$$

where $\{\alpha_{align}, \alpha_{visual}, \alpha_{audio}\}$ are the parameters controlling their relative importance.

4. Main Speaker Dataset

While most of previous speaker datasets [1, 3, 10, 64] are mainly designed for the active speaker detection task, in this work, we further introduce a large-scale dataset for the Main Speaker Detection task. The proposed dataset is collected with three conversation types, i.e. *discussion panel*, *tele-conference*, and *debate*, from several Youtube channels. Particularly, in the discussion panel videos, speakers take turn to speak during the conversation cooperatively. For the second type, videos consist of multiple subject (i.e. 3-5 people) talking through Skype or Zoom in a tele-conference. The third type is more challenging with debate-style videos where there are more interruptions among the two speakers in both cooperative and competitive manners. For each collected video, we select segments of various lengths (i.e. from 6 seconds to 20 seconds) that can represent the property of the corresponding conversation style. In total, the dataset consists of 300 minutes videos. All clips are converted to have 25fps and 16kHz. The bounding box of the main speaker is also annotated.

5. Experimental Results

Data Setting. Our training data include 29 hours of training videos from Lip Reading Sentences 2 (LRS2) [1], and syn-

Table 2. **Main Speaker Audio Separation on LRS2 with and without Domain Alignment**. The higher value is better.

		Ours w/o $\mathcal{L}_{align}$	Ours W $\mathcal{L}_{align}$
SDR (dB)	**1S+N**	14.4	**15.8**
	2S	9.8	**10.3**
	2S+N	7.0	**7.2**
PESQ	**1S+N**	3.1	**3.3**
	2S	2.7	**2.9**
	2S+N	2.5	**2.5**

thetic videos obtaining as presented in Sect. 3.5. The length of synthetic segments varies from 4s to 8s decomposed into 2s short segments. The overlapped ratio of the mixing voice is set to $\frac{1}{3}$. For validation, we adopt the testing set of LRS2, Lip Reading Sentences 3 (LRS3) [3], Columbia [10], and our collected dataset. While LRS2 and LRS3 include 0.5-hour to 1-hour testing videos, Columbia includes an 86-minute panel discussion. We adopt the ground truth for each active speaker in Columbia while annotating bounding boxes using face detection [17] for LRS2 and LRS3.

Audio Data Preprocess. We employ Short Time Fourier Transform (STFT) to represent the audio signal. Our STFT use Han window function which generates magnitude and phase of spectrograms. We set the hop length of 10 ms with a window length of 40ms at a sample rate of 16000Hz.

Visual Data Preprocess. All training videos are re-sampled to a resolution of 160 × 160 pixels at 25 FPS. This chosen resolution results in a feature map composing $N = 6 \times 6$ blocks. During testing phase, we only re-sample the input video to 25 FPS and retain the original resolution.

Network Architectures. We employ 3D VGG-style network for visual deep feature embedding F_v, and 2D VGG-style network for audio embedding F_a. The linear projections $\{\Omega_v^Q, \Omega_v^K, \Omega_v^V, \Omega_a^Q, \Omega_a^K, \Omega_a^V, \Omega^Q, \Omega^K, \Omega^V\}$ are implemented as the fully connected layers that project features to $512 - D$ spaces. The mapping functions $\{\eta_v, \eta_a, \eta\}$ are implemented as residual-style MLP consisting of 2 fully connected layers followed by the normalization layer [8] (the dimension of hidden layers is set to 1024). The audio-visual mask generator D is implemented by a stack of 2 fully connected layers, which predicts both the magnitude mask and the phase mask of the spectrogram. We use the RetinaFace [17] for face detection widely used in face recognition [12, 21, 23, 24, 47, 50, 51, 52, 68].

Model Configurations. Our framework is implemented in PyTorch [58] and all the models are trained on a machine with four NVIDIA P6000 GPUs. The batch size is set to 32 for each GPU. We use RMSProp optimizer with the started learning rate of 0.0001. We set the control parameters to 1.0, i.e $\alpha_{align} = \alpha_{visual} = \alpha_{audio} = 1.0$.

Evaluation Metrics. To compare against prior methods, we adopt four common metrics for localization and audio separation tasks. For single speaker videos, a localization is correct if its lies in the ground-truth bounding box of Main speakers. For multiple-speaker videos, F1 score is adopted for validation. To evaluate Main speaker separation, we adopt the protocol of multi-source speaker audio separation, and estimate the Signal-to-Distortion-Ratio (SDR) [29] and Perceptual Evaluation of Speech Quality (PESQ) [63].

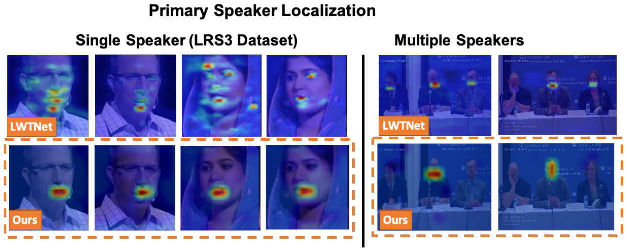


Figure 8. **Main Speaker Localization.** Visualization of attention mask localizing the main speaker. (**Best viewed in color**)

Ablation Study. To study the effectiveness of our proposed cross-domain alignment method, we employ an ablation study with audio separation task on LRS2 using two configurations: without and with $\mathcal{L}_{align}$. We create synthetic testing video samples from LRS2 by combining audios from multiple videos. Three use-cases are evaluated including a primary voice with background noise (1S+N); a primary voice mixed another speaker's voice (2S), and a primary voice mixed another speaker's voice plus background noise (2S+N). We report SDR (dB) and PESQ metrics for these cases in Table 2. By aligning the features of the two domains, the audio-visual correlations can be efficiently extracted and help to consistently improve SDR in all cases.

5.1. Main Speaker Localization

Cooperative Turn-Taking Conversation. In this type, as each speaker takes his/her turn to join the conversation, the main speaker is also the one who is actively speaking during the conversation. The localization accuracy for both single-speaker and multiple-speaker conversations in comparison to previous Active Speaker Detection approaches is reported in Table 3. For each training mode of our model, we also include the configurations that take into account the correlations within and across segments. As can be seen, with the attention mechanisms as well as the domain alignment process, the visual and audio features are better correlated and provide more accurate locations of the main speaker. Moreover, when the spatial-temporal dimension is adopted in configuration (B) and (D), the performance is further boosted. Thanks to the correlations across segments (shown in Fig. 4), the location of each speaker is highly correlated with face of the same subject in other segments and, therefore, enable the tracking consistency of that speaker during the conversation. Our approach outperforms LWT-Net [5] in all datasets with the margins from 0.1% to 4.1%. Fig. 8 shows our localization results compared to LWTNet.

Competitive Turn-Taking Conversation. This type is more challenging as two speakers may speak at the same time. Therefore, although the two speakers can be both active speakers, only one of them is considered as the main speaker while the other one is the interrupter. For this task,

Table 3. **Main Speaker Localization Accuracy with Cooperative Turn-Taking Conversation** (%). For LRS2 and LRS3, a localization is considered correct if it lies within the true bounding box. For Columbia, F1 score is adopted.

	Single Speaker		**Multiple Speakers**					
	LRS2	**LRS3**	**Columbia** (Avg)	**Columbia** (Per subject)				
				Bell	**Boll**	**Lieb**	**Long**	**Sick**
Baseline (Random Pixel)	2.8%	2.9%	8.5%	7.8%	8.6%	9.9%	7.9%	8.7%
Baseline (Center Pixel)	23.9%	25.9%	14.9%	13.0%	11.5%	21.4%	19.2%	17.9%
Multisensory [56]	99.3%	24.8%	52.7%	52.0%	43.8%	62.3%	64.8%	60.9%
Chakravarty et al.[11]	–	–	80.2%	82.9%	65.8%	73.6%	86.9%	81.8%
SyncNet [14]	–	–	89.5%	93.7%	83.4%	86.8%	**97.7%**	86.1%
LWTNet [5]	99.6%	99.7%	90.8%	92.6%	82.4%	88.7%	94.4%	95.9%
(A) **Ours - Block attention**	99.7%	99.8%	92.7%	93.7%	85.0%	87.5%	92.8%	97.2%
(B) + across segments	**99.8%**	**99.9%**	93.4%	95.8%	85.0%	87.5%	92.8%	97.2%
(C) **Ours - Speaker Attention** [†]	100%	100%	93.8%	95.8%	85.0%	91.6%	92.8%	97.2%
(D) + across segments	**100%**	**100%**	**94.9%**	**95.8%**	**88.5%**	**91.6%**	96.4%	**97.2%**

Table 4. **Main Speaker Localization Accuracy in Competitive Turn-Taking Conversation.**

	Discussion Panel	**Tele Conf**	**Debate**
Baseline (Random)	4.8%	3.0%	9.5%
a Baseline (Center)	1.3%	1.1%	5.9%
LWTNet[5]+large_mag	62.8%	55.07%	58.7%
LWTNet[5]+high_corr	88.0%	80.0%	63.3%
Ours	**90.2%**	**83.3%**	**69.4%**

Table 5. **Main Speaker Audio Separation on LRS2.**

	SDR (dB)↑			**PESQ↑**		
	1S+N	**2S**	**2S+N**	**1S+N**	**2S**	**2S+N**
Mix input	1.3	1.31	0.6	1.1	1.1	1.0
SoundOfPixel [72]	9.4	1.5	0.5	1.2	1.1	1.0
Deep-Clustering [38]	9.0	6.0	3.2	2.3	2.3	1.9
Conv-TasNet [49]	–	10.7	–	–	–	–
LWTNet [5]	–	10.8	–	–	3.0	–
Ours (Audio Only)	11.1	9.1	7.0	2.8	2.8	2.5
+ across segments	11.2	9.4	7.1	2.8	2.9	2.6
(A) **Ours - Block Attention**	15.8	10.3	7.2	3.3	2.9	2.5
(B) + across segments	16.6	11.5	8.1	**3.6**	3.1	2.7
(C) **Ours - Speaker Attention**	16.5	10.5	7.5	3.4	3.0	3.0
(D) + across segments	**16.7**	**11.6**	**8.2**	3.4	**3.1**	**3.1**

beside the Random Pixel and Center Pixel baselines, we consider two additional localization strategies. We firstly employ LWTNet [5] to localize all active speakers of each video segment and then choose the main speaker as the one with (1) larger audio magnitude (i.e. *large_mag*), and (2) maximal audio-visual correlation (i.e. *high_corr*). Table 4 reports the localization accuracy on our collected dataset in terms of F1 score against the four baseline approaches. These results again emphasizes the advantages of our proposed approach in the capability of automatically and robustly localize the main speaker in a conversation. The achieved improvements comes from three properties of the proposed model: (1) the present of the contextual attention from both visual and audio domains; (2) the domain feature alignment, and (3) the Cycle Synchronization Loss $\mathcal{L}_{Cyc\_Sync}$ that minimizes the disparity between the localization masks obtained from mixed voices and clean voice.

5.2. Main Speaker Audio Separation

To quantitatively evaluate the capability of audio separation for the proposed approach, we employ the evaluation protocol of [5] and use SDR and PESQ as the validation metrics. Similar to the previous section, we create synthetic testing videos from LRS2 on three cases, i.e. 1S + N, 2S, and 2S + N, and evaluate different configurations of our approach in comparison to previous methods as shown in Table 5. With the spatial-temporal attentions, all configurations that take into account the cross-segment correlations get improvements from 0.3 to 1.2dB of SDR when separating voices of two speakers. Furthermore, the audio-visual attentions also give more cues to improve the separation process. We validate the roles of SROI by adopting two strategies (see Sect. 3.3.1), i.e. block decomposition and face detection. Although the use of face detection can give more focus on face regions and produce further improvements, the block decomposition approach can still attend to the track of the same speaker across segments and give competitive performance. Moreover, our approach with both configurations outperforms LWTNet [5] in SDR and PESQ.

6. Conclusion

This work has presented a novel Audio-Visual Transformer approach for Main Speaker Localization and Audio Separation. Thanks to the introduced attention mechanisms in spatial-temporal dimension together with the domain alignment for better synchronization, our method can effectively localize and highlight the main speaker in both visual and audio channels on multi-speaker conversation videos. Experiments in visual localization and audio separation tasks have shown the advantages of our proposal.

Acknowledgement. This work is supported by NSF Data Science, Data Analytics that are Robust' and Trusted (DART) and Chancellor's Innovation Fund, UAF. It was also funded in part by the Arkansas Biosciences Institute, the agricultural and biomedical research program of the Arkansas Tobacco Settlement Proceeds Act of 2000.

[†] We report the accuracy of the face detection in single-speaker case.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *TPAMI*, page 1–1, 2019.

[2] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv:1809.00496*, 2018.

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. In *INTERSPEECH*, 2019.

[5] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020.

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.

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

[9] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *CVPR*, 2007.

[10] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *ECCV*, 2016.

[11] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. In *ECCV*, 2016.

[12] C. Chen, W. Yang, Y. Wang, K. Ricanek, and K. Luu. Facial feature fusion and model selection for age estimation. In *FG*, 2011.

[13] Joon Son Chung, Bong-Jin Lee, and Icksang Han. Who said that?: Audio-visual speaker diarisation of real-world meetings. In Gernot Kubin and Zdravko Kacic, editors, *INTERSPEECH*, 2019.

[14] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016.

[15] Ross Cutler and Larry Davis. Look who's talking: Speaker detection using video and audio correlation. In *ICME*, 2000.

[16] Pino Cutrone. Profiling performances of l2 listenership: Examining the effects of individual differences in the japanese efl context. *TESOL*, 2019.

[17] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv:1905.00641*, 2019.

[18] Y. Ding, Y. Xu, S. X. Zhang, Y. Cong, and L. Wang. Self-supervised learning for audio-visual speaker diarization. In *ICASSP*, 2020.

[19] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *CVPR*, 2016.

[20] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Deep appearance models: A deep boltzmann machine approach for face modeling. *IJCV*, 2019.

[21] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Ngan Le. Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. *arXiv:1905.10620*, 2019.

[22] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D. Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *CVPR*, 2019.

[23] Chi Nhan Duong, Kha Gia Quach, Ibsa Jalata, Ngan Le, and Khoa Luu. Mobiface: A lightweight deep learning face recognition on mobile devices. In *BTAS*, 2019.

[24] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Hoai Bac Le, and Karl Ricanek Jr. Fine tuning age estimation with global and local facial features. In *ICASSP*, 2011.

[25] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T. Hoang Le, Marios Savvides, and Tien D. Bui. Learning from longitudinal face demonstration–where tractable deep modeling meets inverse reinforcement learning. *IJCV*, 2019.

[26] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T. Hoang Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *ICCV*, 2017.

[27] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *CVPR*, June 2020.

[28] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 2018.

[29] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent. Bss_eval toolbox user guide–revision 2.0. 2005.

[30] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. *NIPS*, 2000.

[31] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through noise: Visually driven speaker separation and enhancement. In *ICASSP*, pages 3051–3055, 2018.

[32] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[33] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, 2019.

[34] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019.

[35] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018.

[36] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[37] J Hershey and JR Movellan. Audio-vision: Locating sounds via audio-visual synchrony. *NIPS*, 1999.

[38] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016.

[39] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019.

[40] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NIPS*, pages 10077–10087, 2020.

[41] Hamid Izadinia, Imran Saleemi, and Mubarak Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *TMM*, 2012.

[42] Zhaozhang Jin and DeLiang Wang. A supervised learning approach to monaural segregation of reverberant speech. *TASLP*, 2009.

[43] Faheem Khan and Ben Milner. Speaker separation using visually-derived binary masks. In *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[44] Naji Khosravan, Shervin Ardeshir, and Rohit Puri. On attention modules for audio-visual synchronization. In *CVPRW*, 2019.

[45] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *CVPR*, 2005.

[46] Bruno Korbar. Co-training of audio and video representations from self-supervised temporal synchronization. 2018.

[47] H. N. Le, K. Seshadri, K. Luu, and M. Savvides. Facial aging and asymmetry decomposition based approaches to identification of twins. *Journal of Pattern Recognition*, 2015.

[48] Qingju Liu, Wenwu Wang, Philip JB Jackson, Mark Barnard, Josef Kittler, and Jonathon Chambers. Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking. *TSP*, 2013.

[49] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *TASLP*, 2019.

[50] K. Luu, T. D. Bui, K. Ricanek Jr., and C. Y. Suen. Age estimation using active appearance models and support vector machine regression. In *BTAS*, 2009.

[51] K. Luu, T. D. Bui, and C. Y. Suen. Kernel spectral regression of perceived age from hybrid facial features. In *FG*, 2011.

[52] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen. Contourlet appearance model for facial age estimation. In *IJCB*, 2011.

[53] Shoji Makino, Te-Won Lee, and Hiroshi Sawada. *Blind speech separation*. Springer, 2007.

[54] Eliya Nachmani, Yossi Adi, and Lior Wolf. Voice separation with an unknown number of multiple speakers. In *Proceedings of the 37th international conference on Machine learning*, 2020.

[55] Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D. Bui, and Khoa Luu. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In *CVPR*, 2021.

[56] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[57] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *IJCV*, 2018.

[58] 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. In *NIPS*, 2019.

[59] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *CVPR*, 2021.

[60] Mohammad H Radfar and Richard M Dansereau. Single-channel speech separation using soft mask filtering. *TASLP*, 2007.

[61] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *WACV*, 2020.

[62] Aarthi M Reddy and Bhiksha Raj. Soft mask methods for single-channel speaker separation. *TASLP*, 2007.

[63] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

[64] 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.

[65] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.

[66] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[67] Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. In *NIPS*. 2019.

[68] Dat T. Truong, Chi Nhan Duong, Khoa Luu, Minh-Triet Tran, and Ngan Le. Domain generalization via universal non-volume preserving approach. In *CRV*, 2020.

[69] Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, Ngan Le, and Khoa Luu. Fast flow reconstruction via robust invertible n × n convolution. *Future Internet*, 2021.

[70] Wenwu Wang, Darren Cosker, Yulia Hicks, S Saneit, and Jonathon Chambers. Video assisted speech source separation. In *ICASSP*, 2005.

[71] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019.

[72] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.

[73] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020.