Audio-Visual Event Localization by Learning Spatial and Semantic Co-Attention

Cheng Xue, Xionghu Zhong, Minjie Cai, *Member, IEEE*, Hao Chen, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—This work aims to temporally localize events that are both audible and visible in video. Previous methods mainly focused on temporal modeling of events with simple fusion of audio and visual features. In natural scenes, a video records not only the events of interest but also ambient acoustic noise and visual background, resulting in redundant information in the raw audio and visual features. Thus, direct fusion of the two features often causes false localization of the events. In this paper, we propose a co-attention model to exploit the spatial and semantic correlations between the audio and visual features, which helps guide the extraction of discriminative features for better event localization. Our assumption is that in an audio-visual event, shared semantic information between audio and visual features exists and can be extracted by attention learning. Specifically, the proposed co-attention model is composed of a co-spatial attention module and a co-semantic attention module that are used to model the spatial and semantic correlations, respectively. The proposed co-attention model can be applied to various event localization tasks, such as cross-modality localization and multimodal event localization. Experiments on the public audio-visual event (AVE) dataset demonstrate that the proposed method achieves state-of-the-art performance by learning spatial and semantic co-attention.**

***Index Terms*—Audio-visual, event localization, cross-modal, co-attention, deep learning.**

I. Introduction

LOCALIZING events in a video is a focused problem in computer science and has wide applications in various domains, such as intelligent surveillance and scene understanding [1], [2]. To tackle the ambiguity and challenges caused by single-modal event localization, the concept of an audio-visual event was recently developed, which defines an event as both audible and visible in a video [3]. In this paper, we tackle the problem of audio-visual event localization with two types of tasks: the *cross-modality localization* (CML) task and the *multimodal event localization* (MMEL) task. The CML task predicts the temporal event boundary in one modality with an input segment of the other modality. The MMEL task predicts event categories for each video segment with both audio and visual signals. Examples of the two tasks are shown in Fig. 1. It should be noted that both CML and MMEL tasks deal with the temporal localization problem. Unless otherwise indicated, the term "localization" refers to "temporal localization" in the following text.

Previous methods [3]–[5] mainly focused on the temporal modeling of audio-visual events with the fusion of audio and visual features and ignored the internal correlations between audio and visual features. Although audio-guided visual attention has been considered to extract visual features, such an attention mechanism is not considered for the audio side. Since much redundant information is contained in a video due to ambient acoustic noise and visual background, it is important to extract discriminative features from both audio and visual modalities for better event localization. In addition, the semantic relationship between audio and visual features has not been well exploited in previous methods. To localize an event that is both audible and visible, it is also important to extract shared semantic information between the two modalities.

This paper aims to localize audio-visual events in a video by developing new strategies for fusing audio and visual information. Inspired by findings in prior psychological and biological studies [6], [7] that the spatial and semantic coordination between hearing and vision is an important mechanism for human perception of the real world, we argue that the modeling of the spatial and semantic correlations between the audio and visual modalities is critical for reliable event localization. For modeling the spatial correlation, while audio features provide spatial attention for encoding discriminative visual features, the encoded visual features in the region of attention can in turn help encode discriminative audio features that are distinct from background noise. For example, in a noisy street, when a person focuses on a driving car, his/her attention to hearing would be to the sound generated by the car compared to other ambient noises. For modeling the semantic correlation, the semantic information encoded by the visual modality and the audio counterpart

Manuscript received 30 August 2020; revised 18 August 2021 and 27 October 2021; accepted 3 November 2021. Date of publication 15 November 2021; date of current version 7 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 61971186 and 61906064, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2020JJ5082, and in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology, and Systems, Beihang University under Grant VRLAB2020B09. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Liangliang Cao. *(Cheng Xue and Xionghu Zhong contributed equally to this article.) (Corresponding author: Minjie Cai.)*

Cheng Xue, Xionghu Zhong, Minjie Cai, and Hao Chen are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: chengxue@hnu.edu.cn; xzhong@hnu.edu.cn; caiminjie@hnu.edu.cn; haochen@hnu.edu.cn).

Wenwu Wang is with the Center for Vision Speech and Signal Processing, Department of Electrical and Electronic Engineering, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3127029.

Digital Object Identifier 10.1109/TMM.2021.3127029

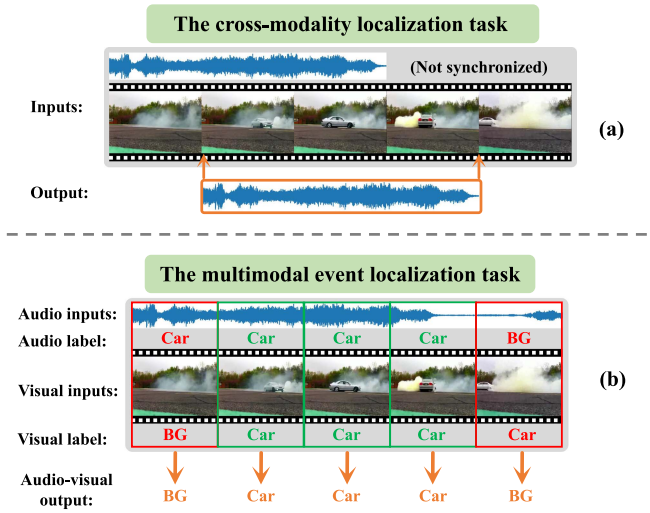


Fig. 1. Illustration of two tasks of the audio-visual event localization problem: *cross-modality localization* (CML) and *multimodal event localization* (MMEL). In the CML task, the temporal boundary of an event in one modality is localized an input segment of the other modality, as shown in (a). In the MMEL task, each video segment is assigned with an event category, as shown in (b). *BG* means background.

should be consistent for an audio-visual event. For example, an event of a car driving happens in a video only when both the appearance and the sound of a car occur.

Technically, we develop an audio-visual event localization framework in which a novel co-attention model is proposed for modeling the mutual correlations between audio and visual modalities. The proposed co-attention model (Section III-B) includes two modules: the *co-spatial attention* (CSPA) module and the *co-semantic attention* (CSEA) module, which work in parallel and model the spatial and semantic correlations between audio and visual features, respectively. Specifically, the CSPA module embeds audio and visual features into a common space to learn spatial attention for visual features, which are then used to learn attention for audio features. The CSEA module augments the shared semantic information within audio and visual features by feature mapping between the two modalities. The output features from the proposed co-attention model effectively exploit the spatial and semantic correlations between the audio and visual modalities. After temporal encoding via recurrent neural networks, the encoded features are applied to both CML and MMEL tasks, as described in Section III-D and III-E. With the proposed co-attention model, our method achieves state-of-the-art performance for both tasks on a public audio-visual event dataset.

In summary, the main contributions of this work include the following:

- A novel co-attention model is proposed to exploit spatial and semantic correlations between audio and visual modalities.
- A unified end-to-end deep framework is developed to solve various tasks of audio-visual event localization.
- The proposed method achieves state-of-the-art performance on various tasks in a public audio-visual event dataset.

II. Related Works

A. Audio-Visual Event Localization

The task of event localization aims to temporally localize events in videos. Previous event localization methods [8], [9] mainly used sound signals to localize whether an event occurred. To tackle the limitations of existing event localization methods based on a single audio or visual modality, the problem of audio-visual event localization [3] was introduced recently, which aims to detect events that are both audible and visible in a single video. Tian *et al.* [3] proposed using audio information to guide visual features for spatial localization. Lin *et al.* [4] used the dual seq2seq method to learn the temporal dependence of audio and visual features. Wu *et al.* [5] proposed using the global event feature as a reference for localizing audio-visual events. Most of the existing audio-visual event localization methods focus on the temporal modeling of audio-visual modalities, and few have studied the mutual relationship between the two modalities.

Different from the previous methods, we propose to jointly model spatial and semantic relationship between the audio and visual modalities for better event localization.

B. Audio-Visual Learning

Previous work on audio-visual learning attempts to fuse or embed audio and visual features [10], [11] and learn the dependencies between audio and visual features [12]–[14] for various tasks such as i) audio-visual representation learning [10]–[12], [15], [16] and ii) audio-visual sound source localization [11], [17]–[23].

1) Audio-Visual Representation Learning: Aytar *et al.* [15] uses the natural synchronicity of audio-visual streams in videos to design a visual teacher network to learn audio representations from unlabeled videos. Arandjelovic and Zisserman [10] introduced an audio-visual correspondence task that learns both audio and visual representations in an unsupervised manner. In addition, several works have temporally sampled audio and visual modalities, thereby using self-supervised learning for audio and visual modalities to learn the temporal correspondence between the two modalities [10]–[12]. Owens and Efros [11] used self-supervised learning to predict whether video frames and audio are temporally aligned.

2) Audio-Visual Sound Source Localization: Zhao *et al.* [21] introduced the PixelPlayer system to take advantage of the natural synchronization of audio and visual modalities and to learn to spatially locate the image region that produced the sound. Senocak *et al.* [22] proposed using audio information to guide sound source localization in visual scenes and built a sound source localization dataset.

Related to the above work, we have also considered how to learn the dependencies between audio and visual features. Specifically, we focus on modeling their spatial and semantic correlations with attention learning for event localization.

C. Attention Learning

The attention mechanism can be understood by comparing it to the human visual system, which selectively focuses on a part of

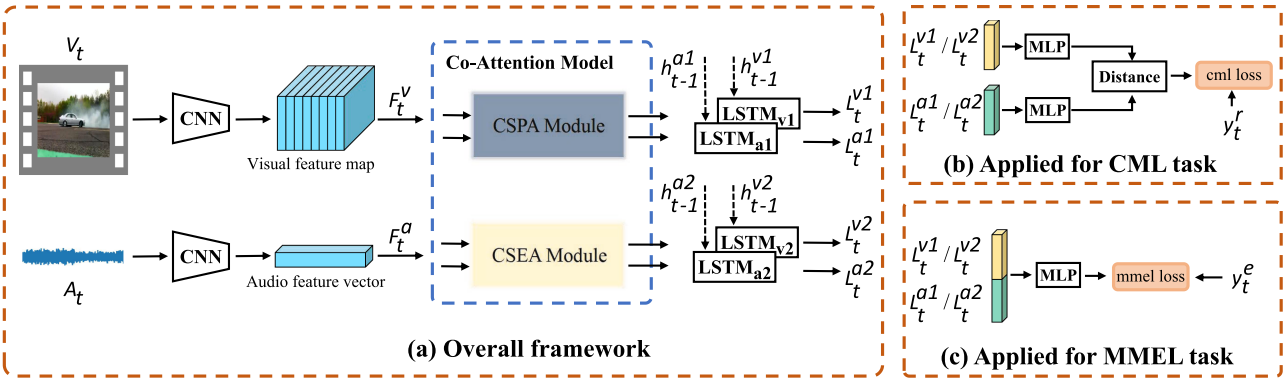


Fig. 2. Illustration of the proposed audio-visual event localization framework. The input is the audio and visual signals in synchronized audio-visual video. The co-attention model is proposed to learn both spatial and semantic correlations between audio and visual features and can be directly applied to solve the CML task (b) and the MMEL tasks (c).

all information while ignoring other visible information [24]. For different tasks, the attention mechanism can learn to focus on the most important part of the task. Recently, attention mechanisms have shown great effectiveness in many areas, such as computer vision [25], audio classification and tagging [26]–[30], natural language processing [31], and visual question answering [32]. Hu *et al.* [33] proposed reweighting semantic features in the feature channel dimension, and Woo *et al.* [34] proposed attention modules on semantic features based on [33]. Chen *et al.* [35] proposed incorporating channel-wise and spatial attention for image captioning. Vaswani *et al.* [31] proposed a self-attention mechanism to solve the problem of machine translation. Lu *et al.* [36] proposed an attention model for visual question answering by jointly modeling images and question attention.

Different from previous works that mainly learned attention for single modal data such as images or texts, we perform attention learning with multimodal data of audio and visual signals. To address the modality gap, we first map the audio and visual features into a common feature space and then exploit the spatial and semantic correlations between the two modalities for the extraction of discriminative features.

III. Method

A. Preliminaries

1) Task Definition: Given a video of T seconds, the audio and the corresponding visual signals are denoted as $(A_t,\ V_t)_{t=0}^{T}$. For the task of cross-modality localization (CML), given a segment of signal from one modality as input, the goal is to predict a vector of $Y^r = [y_1^r, y_2^r, \ldots, y_T^r]$, $y_t^r \in \{0, 1\}$, indicating the event relevance at each segment of the other modality. For the task of multimodal event localization (MMEL), given synchronized audio and visual signals as input, the goal is to predict a vector of event category $Y^e = [y_1^e, y_2^e, \ldots, y_T^e]$ for each segment. The MMEL task has two modes: weakly supervised event localization (WSEL) and supervised event localization (SEL). For SEL, the label of the event category is given for each segment during the training process, while for WSEL, only video-level event labels are provided.

2) Overall Architecture: The overall architecture of the proposed framework is demonstrated in Fig. 2(a). As shown by the information flow, two convolutional neural networks (CNN) are used to extract audio and visual features F_t^a and F_t^v from a segment of audio and visual signals A_t and V_t, respectively. The core component of the framework is a co-attention (CA) model proposed in this work for exploring spatial and semantic correlations between the audio and visual features. Long short-term memory (LSTM) networks are utilized to capture the temporal characteristics of the audio and visual features (L_t^{a1}, L_t^{v1}) and (L_t^{a2}, L_t^{v2}) after the CSPA and CSEA modules, respectively. The CML task in Fig. 2(b) and the MMEL task in Fig. 2(c) are trained separately, with the same input from feature pairs of (L_t^{a1}, L_t^{v1}) and (L_t^{a2}, L_t^{v2}).

B. Co-Attention Model

Since the raw audio features and visual features extracted from CNNs involve noise and redundant information from the background, it is important to focus on discriminative information for better event localization. The co-attention model is proposed to extract discriminative information by exploiting the spatial and semantic correlations between audio and visual features. The co-attention model is composed of a CSPA module and a CSEA module, which are used to model the spatial and semantic correlations, respectively.

1) Co-Spatial Attention Module: Commonly speaking, audio features contain semantic information, that is, the category of events the audio features represent. On the other hand, visual features consist of diverse semantic information in different spatial regions, and only a part of the spatial region is relevant to the event. In this work, we propose a co-spatial attention module to exploit the information in the audio features to guide the extraction of spatially important visual features that are semantically consistent with the audio features.

The network architecture of the co-spatial attention module is shown in Fig. 3(a). Let F_t^a denote the raw audio features with dimensions of C_a. Let F_t^v denote raw visual features with dimensions of $[W, H, C_v]$, where (W, H) denotes the spatial dimension and C_v denotes the number of channels. The raw audio

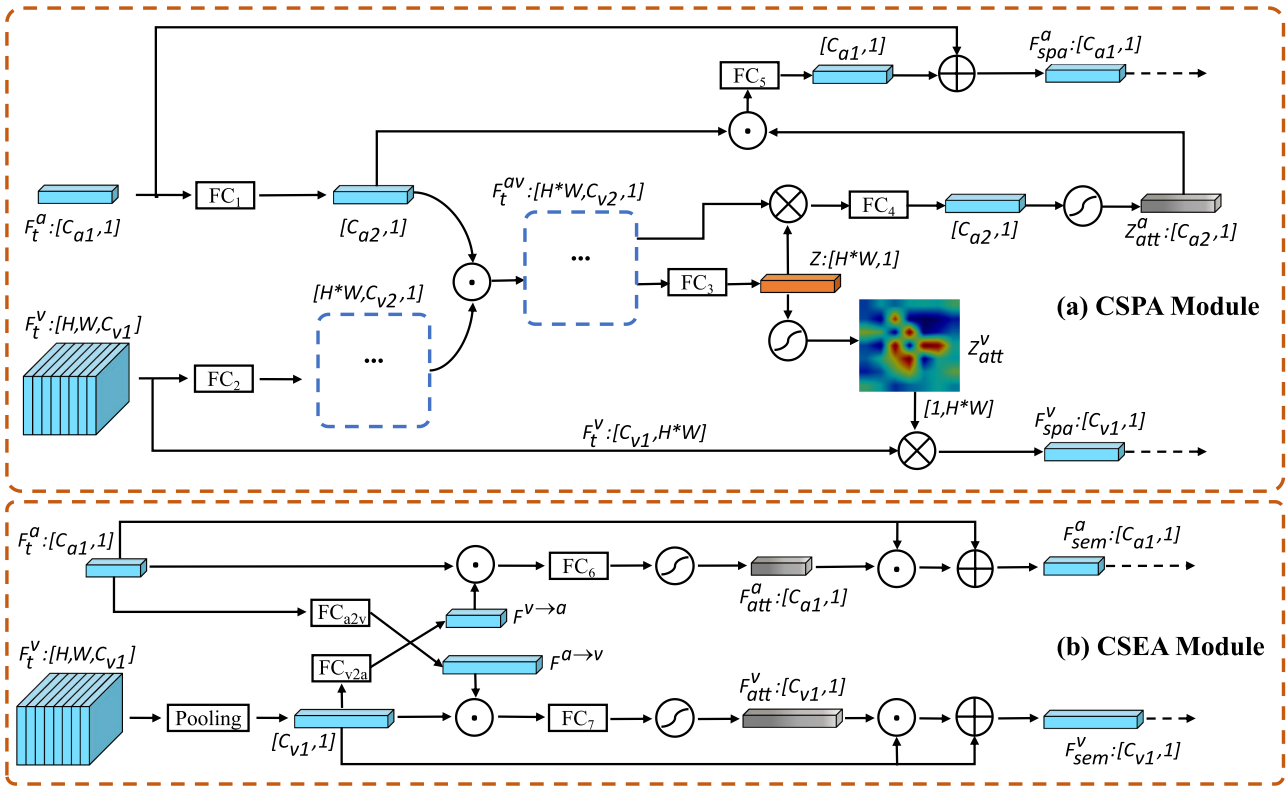


Fig. 3. The proposed co-attention model includes two modules, CSPA and CSEA. The CSPA module is proposed to learn the spatial relationship between audio and visual features, as shown in (a). The CSEA module is developed to model the semantic relationship between audio and visual features, as shown in (b). In the figure, (H, W) represents the spatial dimension, and $C_{vi/ai}$ represents the channel (or semantic) dimension at different stages, where $C_{v2} = C_{a2} = H \times W$. ' $\otimes$ ' denotes matrix multiplication, ' $\odot$ ' denotes Hadamard elementwise product, and '$\oplus$ ' denotes elementwise addition.

and visual features are first embedded into a common feature space from which spatial attention is estimated as follows:

$$\text{fusion} \begin{cases} F_t^{av} = f_1\left(F_t^a\right) \odot f_2\left(F_t^v\right) \\ Z = f_3\left(F_t^{av}\right) \\ Z_{att}^v = softmax(Z) \end{cases} \tag{1}$$

where ' $\odot$ ' denotes the Hadamard elementwise product, $f_i(\cdot)$ represents the i-th feature transformation component implemented with a fully connected (FC) neural network and is shown as FC_i in Fig. 3, and Z_{att}^v represents the spatial attention for visual features.

The goal of this module is to extract discriminative audio and visual features with spatial attention. To extract discriminative visual features, we use Z_{att}^v as spatial attention to filter the raw visual features in the spatial dimension. For extracting discriminative audio features that are correlated with the visual features, we first encode Z to learn an attention vector and then perform attention operations on the raw audio features. The procedure of extracting audio and visual features with spatial attention is presented as follows:

$$\begin{aligned} &\text{For } F^a: \begin{cases} Z_{att}^a = softmax\left(f_4\left(F_t^{av} \otimes Z\right)\right) \\ F_{spa}^a = f_5\left(Z_{att}^a \odot f_1\left(F_t^a\right)\right) + F_t^a \end{cases} \\ &\text{For } F^v: \quad F_{spa}^v = F_t^v \otimes \left(Z_{att}^v\right)^T \end{aligned} \tag{2}$$

where ' $\otimes$ ' denotes matrix multiplication, F_{spa}^a and F_{spa}^v represent audio and visual features after the CSPA module.

2) Co-Semantic Attention Module: In audio-visual event localization, it is important to ensure that the semantic information represented by the audio and visual features corresponds to the same event; otherwise, no event would be detected. Although the CSPA module helps extract discriminative audio and visual features with spatial attention, it cannot guarantee semantic consistency between the two features. Therefore, the co-semantic attention module is designed based on the principle that the semantic information in audio and visual features of the same event should match each other.

The CSEA module is shown in Fig. 3(b). The network design is inspired by prior work in biology [7], [37], which has shown that the human perception system can respond to the visual picture of an event by hearing the sound of the event, and vice versa. As visual features F_t^v contain both spatial and semantic information, we use a global average pooling operation on the spatial dimension of F_t^v to obtain its semantic information. To obtain the semantic correspondence between the audio and visual features, we first map features of one modality into the feature space of the other modality. Then, semantic correspondence (attention) can be obtained by cross-checking the features of two modalities in the common feature space of each modality. Finally, semantic attention is applied to the audio and visual features to enhance the part of features that corresponds to the

same semantic information. The procedure is as follows:

$$\begin{aligned} &i) \quad F^{a\to v} = f_{a2v}\left(F_t^a\right),\ F^{v\to a} = f_{v2a}\left(G(F_t^v)\right) \\ &ii) \quad \begin{cases} F_{att}^a = softmax(f_6(F_t^a \odot F^{v\to a})) \\ F_{att}^v = softmax(f_7(F_t^v \odot F^{a\to v})) \end{cases} \\ &iii) \quad F_{sem}^a = (F_{att}^a + 1) \odot F_t^a,\ F_{sem}^v = (F_{att}^v + 1) \odot F_t^v \end{aligned} \tag{3}$$

where F_{sem}^a and F_{sem}^v represent audio and visual features obtained after the CSEA module, G represents the global average pooling operation, and f_{a2v}, f_{v2a} represents the feature mapping implemented with fully connected networks.

C. Temporal Modeling

To capture the temporal characteristics of audio-visual events, we use LSTM networks to temporally encode the output features of the co-attention model, which is commonly incorporated in other audio-visual event localization methods. The temporal modeling process takes as input the output features of the CSPA and CSEA modules separately. Since the temporal modeling processing for the two modules is similar, we describe only the module with the CSPA stream for simplicity. The process is written as follows:

$$\begin{aligned} L_t^{a1}, \quad \left(h_t^{a1}, c_t^{a1}\right) &= LSTM_{a1}\left(F_{spa}^a, \quad \left(h_{t-1}^{a1}, c_{t-1}^{a1}\right)\right) \\ L_t^{v1}, \quad \left(h_t^{v1}, c_t^{v1}\right) &= LSTM_{v1}\left(F_{spa}^v, \quad \left(h_{t-1}^{v1}, c_{t-1}^{v1}\right)\right) \end{aligned} \tag{4}$$

where L_t^{a1}/L_t^{v1} represent the output of temporal audio and visual features, h_t^{a1}/h_t^{v1} represent the hidden state vectors, and c_t^{a1}/c_t^{v1} represent the memory cell state vectors of $LSTM_{a1}$ and $LSTM_{v1}$ at time step t, respectively.

Similar to the CSPA module, temporal audio and visual features L_t^{a2}/L_t^{v2} of the CSEA module are obtained as the output of $LSTM_{a2}$ and $LSTM_{v2}$.

D. Cross-Modality Localization

Given an audio or visual segment, the CML task aims to match the corresponding visual segment (A2V) or audio segment (V2A) of the same event. In this section, we introduce the procedure of how to use the output features from LSTMs to complete the CML task. Since the procedures of V2A and A2V are similar, we only introduce the CML task in the V2A mode for simplicity.

In the training phase, given the CSPA stream as an example, let L_t^{a1} and L_t^{v1} denote the audio and visual features output from LSTM_{a1} and LSTM_{v1} at time t in the CSPA stream. Two multilayer perceptrons (MLPs), as shown in Fig. 2(b), each containing two fully connected layers with output sizes 128 and 64, respectively, are used to encode L_t^{a1} and L_t^{v1} into a common feature space as $\mathbf{m}_t^{a1}$ and $\mathbf{m}_t^{v1}$ for matching. With event relevance label $y_t^r \in \{0, 1\}$ for each video segment, our goal is to learn the model parameters by minimizing the mismatch between the corresponding audio and visual segments. We use the contrastive loss [38] for training, as formulated below:

$$\begin{aligned} \mathcal{L}_t^{C1} = &\ y_t^r S_\theta^2\left(\mathbf{m}_t^{v1}, \mathbf{m}_t^{a1}\right) \\ &+ \left(1 - y_t^r\right)\left(\max\left(0, \tau - S_\theta\left(\mathbf{m}_t^{v1}, \mathbf{m}_t^{a1}\right)\right)\right)^2 \end{aligned} \tag{5}$$

where $\mathcal{L}_t^{C1}$ denotes the contrastive loss for a pair of segments in the calculation stream of the CSPA module, S_θ denotes the Euclidean distance with model parameter θ, and $\tau = 2.0$ is the threshold for contrastive loss. Similar to the CSPA stream, we can obtain the contrastive loss $\mathcal{L}_t^{C2}$ in the calculation stream of the CSEA module. Overall, the total loss for the CML task is $\mathcal{L}_t^C = \lambda_c \mathcal{L}_t^{C1} + (1 - \lambda_c)\mathcal{L}_t^{C2}$, where $\lambda_c \in (0, 1)$ is a hyperparameter balancing the contribution of the CSPA and CSEA streams. We evaluate the impact of various values of λ_c in Section IV-C.

In the inference phase, a visual segment V_r is chosen from $\{V_i\}_{i=0}^T$, and then a best-matching audio segment is obtained from $\{A_i\}_{i=0}^T$ by calculating the Euclidean distance for each pair of segments in a sliding window manner. The pairwise distance is computed from the CSPA and CSEA streams separately, and the average distance of the two streams is used.

E. Multimodal Event Localization

The goal of multimodal event localization is to predict event categories for each video segment given audio and visual features. We consider both supervised event localization (SEL) and weakly supervised event localization (WSEL). In the SEL setting, event category label $\mathbf{y}_t^e$ of a one-hot vector is provided for each segment, and the output is a segment-level classification vector $\mathbf{e}_t$, both with the shape of [1, C], where $C = 29$ represents the number of event categories. In the WSEL setting, the event category label is video-level and represented as $\mathbf{y}^w$, indicating the occurrence of an event for the whole video. The output is a video-level classification vector $\mathbf{v}$ obtained by global max-pooling along the temporal dimension of all segment-level classification vectors $\{\mathbf{e}_t\}$, with its element computed as $v[i] = \max_{t=1,2,\ldots,T} e_t[i]$.

At training time, given the CSPA stream for example, the concatenation of L_t^{a1} and L_t^{v1} is first encoded by a MLP as shown in Fig. 2(c), containing two fully connected layers with output sizes 64 and 29, respectively, and then the event classification score $\mathbf{e}_t$ is obtained for each segment. We use the multilabel soft-margin loss for training:

$$\begin{aligned} \mathcal{L}^{S1} = &-\frac{1}{C} * \sum_{t=0}^{T} \sum_{j=0}^{C-1} \mathbf{y}_t^e[j] * \log\left(\frac{\exp(\mathbf{e}_t[j])}{1 + \exp(\mathbf{e}_t[j])}\right) \\ &+ \left(1 - \mathbf{y}_t^e[j]\right) * \log\left(\frac{1}{1 + \exp(\mathbf{e}_t[j])}\right) \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{L}^{W1} = &-\frac{1}{C} * \sum_{j=0}^{C-1} \mathbf{y}^w[j] * \log\left(\frac{\exp(\mathbf{v}[j])}{1 + \exp(\mathbf{v}[j])}\right) \\ &+ \left(1 - \mathbf{y}^w[j]\right) * \log\left(\frac{1}{1 + \exp(\mathbf{v}[j])}\right) \end{aligned} \tag{7}$$

where $\mathcal{L}^{S1}$ and $\mathcal{L}^{W1}$ represent the loss functions of the SEL and WSEL settings, respectively. In a similar way, we can obtain the SEL loss $\mathcal{L}^{S2}$ and the WSEL loss $\mathcal{L}^{W2}$ in the CSEA stream. Overall, the total losses for SEL and WSEL are $\mathcal{L}^S = \lambda_s \mathcal{L}^{S1} + (1 - \lambda_s)\mathcal{L}^{S2}$ and $\mathcal{L}^W = \lambda_w \mathcal{L}^{W1} + (1 -$

$\lambda_w)\mathcal{L}^{W2}$, respectively. λ_s and λ_w are hyperparameters balancing the contribution of the CSPA and CSEA modules for the two subtasks respectively.

At inference time, the model averages the outputs of the CSPA and CSEA streams as final prediction scores of audio-visual events.

IV. Experiments

A. Experiment Setup

Dataset: We conduct experiments on a large public AVE dataset [3]. The AVE dataset consists of 4143 videos, each lasting 10 seconds. It is constructed by Tian *et al.* [3] as a subset of Audioset [39] and ensures that only one audio-visual event exists in each video segment. The dataset is split into training, validation and test sets with 3339, 402, and 402 videos, respectively. For cross-modality localization, the AVE dataset provides a label $y_t^r \in \{0, 1\}$ for each one second video segment, where 1 represents the audio and visual signals of the same segment match each other (i.e., an audio-visual event happens in the segment) and 0 otherwise. For supervised event localization, annotations of event categories are provided for each one-second video segment in both the training and testing stages, with a total of 28 categories (e.g., horse, violin, mandolin, helicopter, baby cry, etc.) plus 1 background. For weakly supervised event localization, only video-level annotations of event categories are provided in the training stage. However, it is evaluated at the segment-level in the testing stage, similar to supervised event localization.

Evaluation Metrics: Accuracy is used as the evaluation metric for both tasks, following the evaluation routing in previous work [3].

For the CML task, accuracy is computed as the percentage of correct matching over all matched segment pairs from the test videos. A matched segment pair is composed of a query segment in the source modality and a best-matching segment in the target modality. A matched segment pair is considered correct if the two segments are temporally aligned. For a video with $T = 10$ segments, the accuracy by chance is 10%.

For the MMEL task, accuracy is computed as the percentage of correct classification over all test segments for both supervised and weakly supervised settings. In the dataset, the background class has taken up the highest percentage of 17.1%, which implies accuracy by chance.

Implementation Details: For feature extraction before the proposed co-attention model, the duration of each video in the dataset is $T = 10$ s, and we extract audio and visual features for each one-second video segment. For visual input, the original frame rate in the dataset is not fixed and is greater than 16. For consistency, 16 images are sampled in each one-second segment, and visual features are extracted through a VGG19 network [40] pretrained on Imagenet [41]. For audio input, the original audio signal has a sampling frequency of 16 kHz. Short-time Fourier-transform (STFT) is performed for each 25 ms window with a step of 10 ms. The STFT signal in the frequency domain is then transformed into mel-scale filter banks with 64 bins. This gives a mel-spectrogram "image" with dimensions of 96×64 for each one-second segment. Audio features are extracted by feeding the mel-spectrogram image into a VGGish network [42] pretrained on Audioset [39]. For more details, please refer to [42]. Note that the VGG16 and VGGish networks correspond to the CNNs in Fig. 2(a).

TABLE I
Implementation Details of Each Fully Connected Layer in the Network

	hidden layers	output of each layer
FC_1	2	[512,49]
FC_2	2	[512,49]
FC_3	1	[1]
FC_4	2	[256,49]
FC_5	1	[128]
FC_{v2a}	1	[128]
FC_{a2v}	1	[512]
FC_6	1	[128]
FC_7	1	[512]

In the co-attention model, the values of C_{a1}, C_{v1}, C_{v2} and C_{a2} are set as 128, 512, 49 and 49, respectively. The spatial dimension $H \times W$ of visual features is 7×7. The implementation details of each fully connected layer in the network are shown in Table I. For temporal modeling, we use bidirectional LSTMs (BLSTMs) for audio and visual features, where the number of hidden states and output are 128 and 256, respectively, for both modalities. The hyperparameters for balancing the CSPA and CSEA modules are empirically set as $\lambda_c = 0.6$, $\lambda_s = 0.3$, and $\lambda_w = 0.7$ for the CML, SEL and WSEL tasks.

In the proposed method, both the pretrained VGG19 and VGGish networks are fixed in the CML and MMEL tasks. The co-attention model and the LSTM network are first trained for the MMEL task. Then, for the CML task, we initialize the co-attention model and the LSTM network with parameters learned from the MMEL task. We found that this can help accelerate the convergence of training for the CML task.

The network is optimized using the Adam optimizer with an initial learning rate of 0.001. In the MMEL task, the model is trained for 300 epochs in total with a batch size of 64, and in the CML task, the model is trained for 30 epochs in total with a batch size of 32.

B. Comparison With State-of-The-Art

In this section, we compare our method with state-of-the-art methods for both tasks of cross-modality localization and multimodal event localization. The compared methods are as follows:

DCCA [43]. By combining deep neural network and canonical correlation analysis (CCA), it tried to learn the feature mapping of two views that are maximally correlated at the same time.

ED-TCN [44]. It introduced a temporal model of temporal convolutional networks (TCNs) to capture long-range patterns by using a hierarchy of temporal convolutional filters.

Audio-visual w /att and **AVDLN** [3]. It proposed an audio-guided visual attention mechanism. In addition, it also proposed a dual multimodal residual network (DMRN) to fuse the information over the two modalities.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CROSS-MODALITY LOCALIZATION TASK. ACCURACY (%) IS USED AS THE EVALUATION METRIC

Method	A2V	V2A	Average
DCCA [43]	34.8	34.1	34.5
AVDLN [3]	44.8	35.6	40.2
DAM [5]	47.1	48.5	47.8
Ours	**49.0**	**51.0**	**50.0**

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE SUPERVISED SETTING OF MULTIMODAL EVENT LOCALIZATION TASK. † INDICATES THAT THE METHOD WAS RE-IMPLEMENTED WITH THE SAME PRE-TRAINED VGG-19 FEATURE FOR A FAIR COMPARISON

Method	Accuracy (%)
ED-TCN [44]	46.9
Audio-visual [3]	71.4
AVSDN† [4]	72.6
Audio-visual+Att [3]	72.7
DAM [5]	74.5
Ours	**76.5**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE WEAKLY SUPERVISED SETTING OF MULTIMODAL EVENT LOCALIZATION TASK

Method	Accuracy (%)
Audio-visual [3]	63.7
AVSDN† [4]	63.6
Audio-visual+Att [3]	66.7
Ours	**70.2**

AVSDN [4]. It is based on sequence-to-sequence and autoencoders and exploits global and local event information in a seq2seq [45] manner.

DAM [5]. It proposed a dual attention matching (DAM) module to capture the global information in a long time and the local temporal information through a global cross-check mechanism.

1) Cross-Modality Localization: Table II shows the performance comparison between our method and the current state-of-the-art methods on the cross-modality localization task. DCCA [43] does not work well on the task, probably since it does not consider the specific relationship between the audio and visual features. Although AVDLN [3] considers audio-guided spatial attention for visual features, it does not model their semantic relationship. DAM [5] outperforms previous methods by encoding the global event information as a reference when localizing audio-visual events. Our method achieves the highest average accuracy of 50.0% by modeling the spatial and semantic relationship between audio and visual features. Specifically, on the V2A task, our method improves the accuracy from 48.5% to 51.0%.

2) Multimodal Event Localization: We also evaluate our method on the MMEL task, which aims to predict the event category for each video segment. The performance comparison with state-of-the-art methods for multimodal event localization in supervised and weakly supervised settings is shown in Table III and Table IV, respectively. From Table III, it can be seen that ED-TCN [44], a state-of-the-art temporal action labeling method, is not suitable for the MMEL task. AVSDN [4] proposed using the Seq2Seq mechanism to capture the temporal dependencies of audio and visual features for event prediction. Tian *et al.* [3] considered the audio-guided spatial attention mechanism for visual features to improve the performance. DAM [5] reported higher accuracy in the supervised setting by considering global information on localizing short-term events. Similar results can be seen in Table IV. Our method outperforms previous methods on both supervised (76.5%) and weakly supervised settings (70.2%).

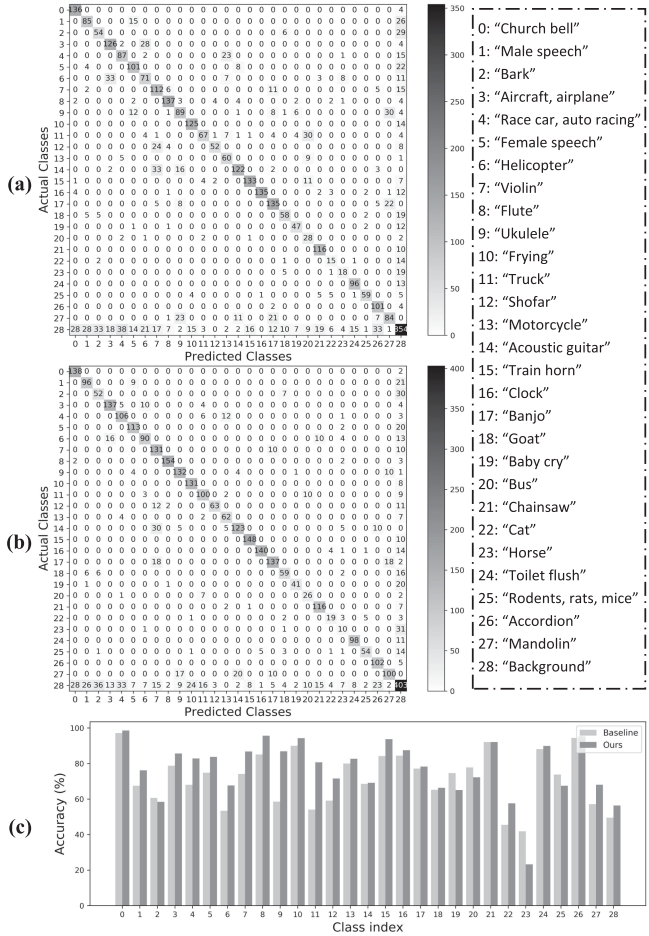


Fig. 4. The confusion matrix of the baseline [3] and the proposed method on the MMEL task are shown in (a) and (b), respectively. (c) compares the accuracy of different event classes between the baseline and the proposed method. The indexes of event classes are shown within the dotted box.

To demonstrate how our method works for different event categories, we show the confusion matrix of our method on the MMEL task in Fig. 4. Our method outperforms the baseline method in most events, with several events largely improved by our method, such as index 4: "Race car," index 11: "Truck," and index 15: "Train horn". These events have sufficient training samples with unique correspondence between visual and audio signals and therefore can be well modeled by our method. However, for some events with relatively small data proportions, such as index 19: "Baby cry" and index 23: "Horse," the proposed method performs slightly worse than the baseline. The reason might be that the data samples of these events are not sufficient

TABLE V
ABLATION STUDY OF DIFFERENT MODULES OF THE PROPOSED CO-ATTENTION MODEL ON FOUR SUB-TASKS. ACCURACY (%) IS USED AS THE EVALUATION METRIC

Method	A2V	V2A	SEL	WSEL
CSPA	39.3	33.3	74.1	68.0
CSEA	48.5	50.7	72.9	66.4
(CSPA /aa) + CSEA	48.5	49.8	75.5	69.8
CSPA + (CSEA /v2a)	48.5	48.0	75.9	70.1
CSPA + (CSEA /a2v)	44.0	47.3	76.3	70.1
CSPA + CSEA	**49.0**	**51.0**	**76.5**	**70.2**

for the proposed method to effectively model the correlations between the two modalities in these events.

C. Ablation Studies

To examine how different parts of our proposed co-attention model contribute to the final performance on the four tasks, we conducted an ablation study by removing a subset of our full model. For all baselines, LSTMs are adopted for temporal modeling. Details of different baselines are explained as follows:

Single attention module. We examine the contribution of a single CSPA module or CSEA module and denote them as **CSPA** and **CSEA**, respectively.

CSPA module without audio attention. To evaluate the contribution of the proposed visual-guided audio attention in the CSPA module, we remove this part and test the remaining model and denote this baseline as **(CSPA /aa)+CSEA**.

CSEA module with one branch. To examine the contribution of semantic attention from either the audio or visual branch, we remove the attention calculated with mapped audio features $F^{v\to a}$ or mapped visual features $F^{a\to v}$, respectively. These two baselines are denoted as **CSPA+(CSEA /v2a)** and **CSPA+(CSEA /a2v)**.

The ablation study results are shown in Table V. Comparing single attention modules, the performance of the CSEA module is significantly better than that of the CSPA module in the CML task (A2V and V2A). This shows that the CSEA module can effectively learn the semantic relationship between audio and visual features so that it can accurately match audio- and visual-related segments in the CML task. However, in the MMEL task (SEL and WSEL), the performance of the CSPA module is better than that of the CSEA module, which implies that the spatially augmented audio and visual features by the CSPA module are more discriminative for predicting the category of audio-visual events. Comparing (CSPA /aa)+CSEA with our full model (CSPA+CSEA), it can be seen that removing the proposed visual-guided audio attention in the CSPA module would degrade the performance. It is also interesting to note that the performance degradation is more severe by removing audio-to-visual mapping (CSPA+CSEA /a2v) than by removing visual-to-audio mapping (CSPA+CSEA /v2a) in the CML task, which indicates that mapping from audio to visual features is more critical in cross-modal mapping. Most importantly, the combination of CSPA and CSEA achieves the best performance in all tasks.

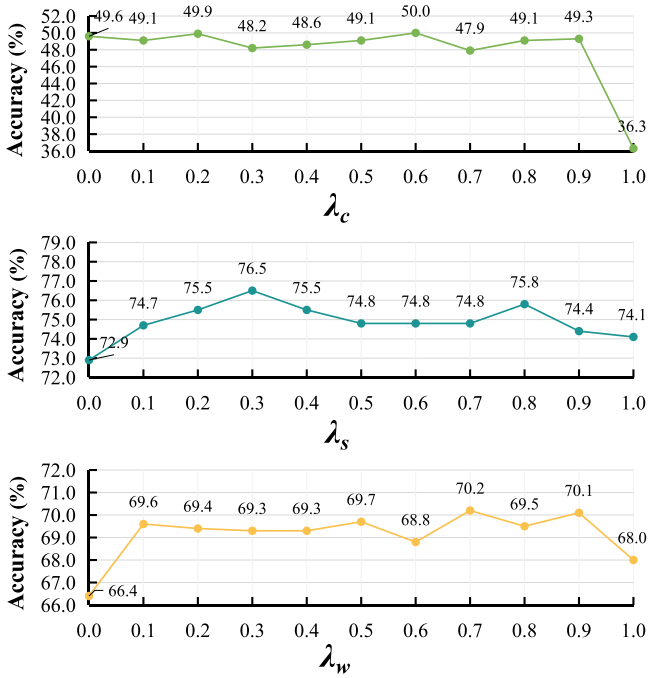


Fig. 5. Illustrations of the effect of hyperparameters on the performance of the proposed method. λ_c, λ_s and λ_w are used to balance the CSPA stream and the CSEA stream in the CML, SEL, and WSEL tasks, respectively.

In Sections III-D and III-E, we introduce three separate hyperparameters λ_c, λ_s and λ_w to balance the contribution of the CSPA and CSEA modules in various tasks. To examine the influence of these three hyperparameters on the performance of our method, we conducted experiments with different values of hyperparameters and the results are shown in Fig. 5. By studying the performance variation with λ_c for the CML task, we can see that the performance would significantly decrease when the CSPA module is used alone ($\lambda_c = 1.0$), which is consistent with the results in Table V. By studying the performance variation with λ_s and λ_w for the MMEL task (including SEL and WSEL), we can see that the performance would decrease when either the single CSPA module ($\lambda_s = 1$ and $\lambda_w = 1$) or the CSEA module is used alone ($\lambda_s = 0$ and $\lambda_w = 0$), which indicates that both the CSPA and CSEA modules are needed for audio-visual event localization. Moreover, when the two attention modules are both used, the change in the relative contribution of the two modules ($\lambda \in [0.1, 0.9]$) only slightly affects the performance. This demonstrates not only the superior performance but also the robustness of our proposed co-attention model.

D. Visualization and Qualitative Results

1) Attention Visualization: The CSPA module is proposed to extract discriminative audio and visual features by exploring spatial attention. Here, we visualize the spatial attention of visual features in Fig. 6 to show the CSPA module's ability to learn meaningful spatial relationships between audio and visual modalities. It can be seen from the figure that as the event evolves with time, our method can always capture the spatial position

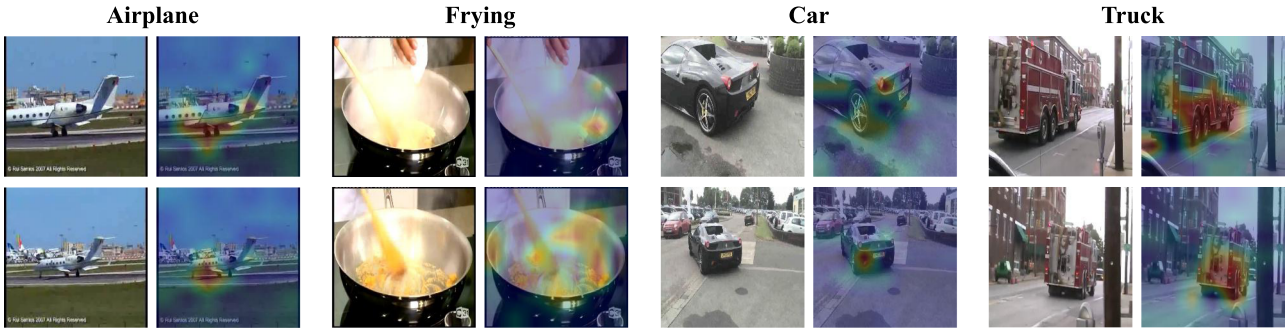


Fig. 6. Visualization of spatial attention learned by the CSPA module on four example events. The figure above visualizes the attention area of the four audio-visual events in the visual space. From top to bottom, it represents two segments selected in the time sequence in the audio-visual video sequence.

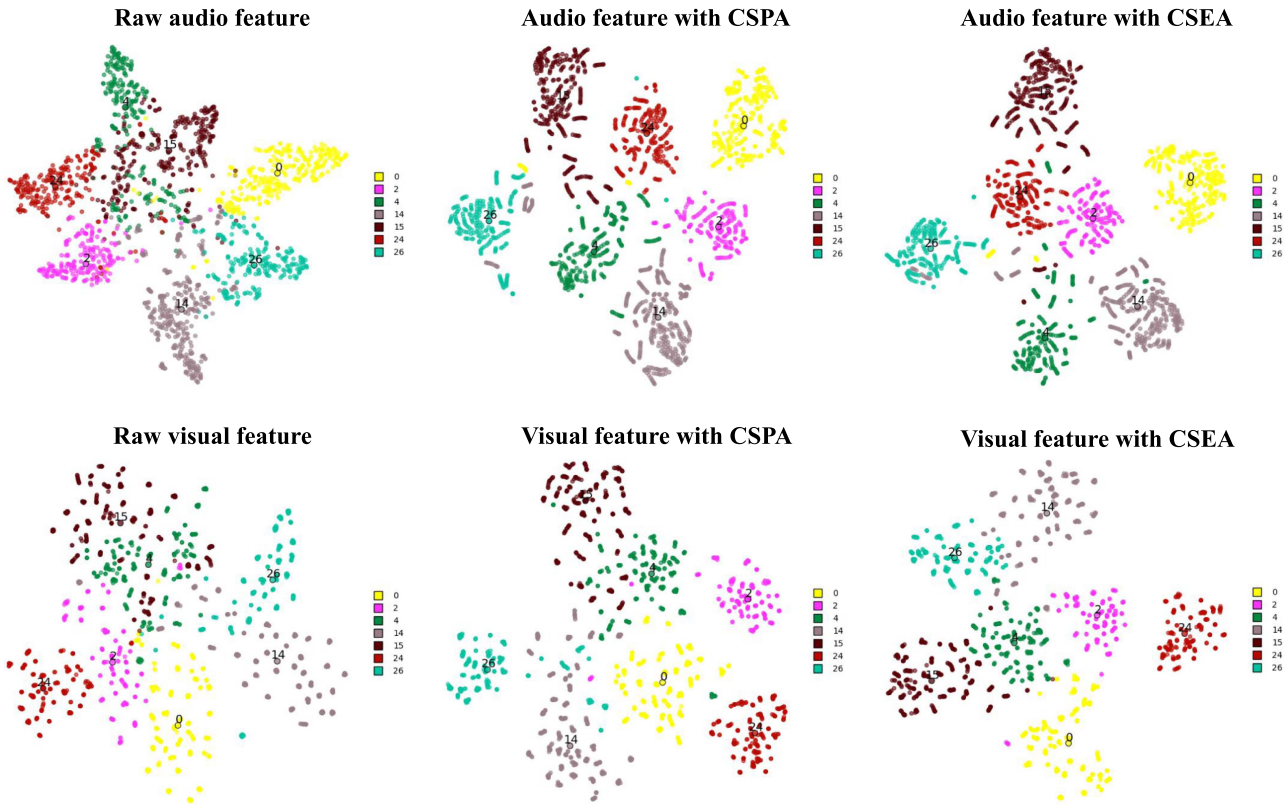


Fig. 7. Illustration of the feature learning of the CSPA stream and CSEA stream in our proposed method. The number indices of 0, 2, 4, 14, 15, 24 and 26 represent "Church bell," "Bark," "Race car, auto racing," "Acoustic guitar," "Train horn," "Toilet flush" and "Accordion", respectively.

where the event occurs. For example, the truck in the 4th example of the figure moves from left to right, and the region of the truck remains highlighted by spatial attention.

2) Feature Learning Visualization: The goal of our co-attention model is to remove redundant information and extract discriminative features for different events by exploiting correlations between audio and visual modalities. To verify that our model works as expected, we use a high-dimensional data visualization tool of t-SNE [46] to visualize the features of different events before and after our co-attention model. We chose several classes with high proportions among all classes for better illustration, and the visualization is shown in Fig. 7. The audio (as well as visual) features of different events with either the CSPA or CSEA modules are separated more apart from each other than raw features. Moreover, the features with CSPA or CSEA modules show certain complementary properties. Taking audio features as an example, while "Race car, auto racing" (index 4) and "Train horn" (index 15) are close in features with CSPA, they are apart in features with CSEA. This partly explains why the combination of CSPA and CSEA works better than the individual module.

3) Qualitative Results: Fig. 8 shows qualitative results on the CML (with two manners of V2A and A2V) and MMEL tasks. For V2A, given a visual segment as input, the audio segment corresponding to the same event is localized for synchronization. A correct example is shown in Fig. 8(a) in which our method correctly matches the visual segment of the baby crying with the corresponding audio segment. A failure case in A2V is shown in

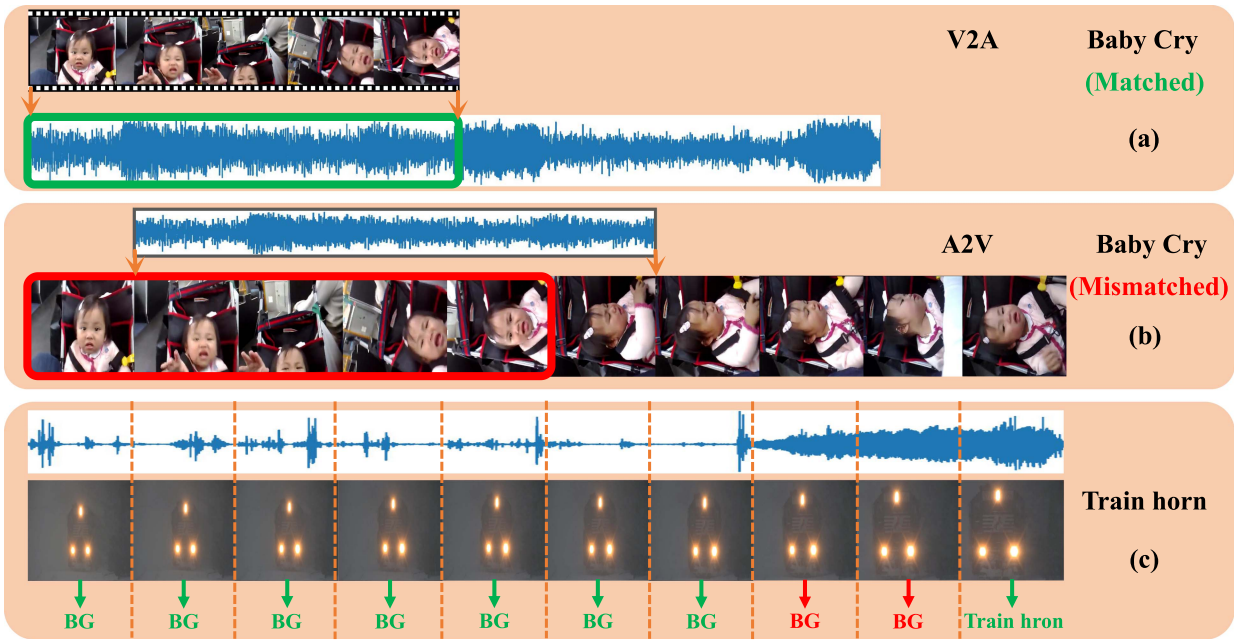


Fig. 8. Qualitative results. The time dimension of the audio-visual sequence is shown from the left to the right of the figure. Green indicates the correct result, red indicates the wrong result, and *BG* denotes the background event.

Fig. 8(b). Our method could not correctly localize the visual segment of the baby crying with the corresponding audio segment. In such failure cases, mismatch is often caused by the inability to correctly identify the event category of the input segment. The qualitative result of the multimodal event localization task is shown in Fig. 8(c), where the event categories in green and red indicate correct and incorrect classifications, respectively. Although our method correctly predicted in most video segments, the eighth and ninth segments, where a train horn event happens, are incorrectly predicted as a background event. In this example, although the audio signal is discriminative, the visual difference between the train horn event and the background event is ambiguous. To tackle this challenge, event-dependent importance weights of different modalities might be useful.

V. Conclusion and Future Work

In this work, we proposed a new end-to-end deep framework for the audio-visual event localization problem. Based on the assumption that the audio and visual features share common semantic information in an audio-visual event, we propose the co-attention model to exploit the spatial and semantic correlations between the audio and visual modalities through attention learning. Specifically, the co-attention model includes the CSPA and CSEA modules to model the spatial and semantic relationship between the two modalities, respectively. The experimental results show that 1) our proposed co-attention model can extract discriminative audio and visual features by fully exploiting the spatial and semantic correlations between the audio and visual modalities. 2) In both the CML and MMEL tasks, our proposed method performs significantly better than previous methods.

Our current model exploits correlations between audio and visual modalities based on an assumption that useful information of an event in the two modalities occurs simultaneously in synchronized data. Therefore, it might be limited to tackle a more challenging case of time-delay events, such as "thunder and lightning," where useful information does not occur simultaneously even though the two modalities of a video are synchronized. To address this case, we consider incorporating temporal modeling within our co-attention model in a unified way, which is left as our future work.

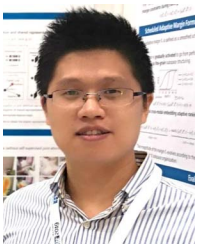
References

[1] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, pp. 257–267, 2007.
[2] C. Canton-Ferrer *et al.*, "Audiovisual event detection towards scene understanding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2009, pp. 81–88.
[3] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 247–263.
[4] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality Seq2Seq network for audio-visual event localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 2002–2006.
[5] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6292–6300.
[6] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychol.*, vol. 5, no. 1, pp. 1–29, 1993.
[7] R. Q. Quiroga, A. Kraskov, C. Koch, and I. Fried, "Explicit encoding of multimodal percepts by single neurons in the human brain," *Curr. Biol.*, vol. 19, no. 15, pp. 1308–1313, 2009.
[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, pp. 1–13, 2013.
[9] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6440–6444.

[10] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 609–617.
[11] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.
[12] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 435–451.
[13] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," 2017, *arXiv:1706.00932*.
[14] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 71–88.
[15] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 892–900.
[16] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7763–7774.
[17] J. R. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 813–819.
[18] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.
[19] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, pp. 186–200, 2015.
[20] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.
[21] H. Zhao *et al.*, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 570–586.
[22] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4358–4366.
[23] Y. Liu, V. Kilic, J. Guan, and W. Wang, "Audio-visual particle flow SMC-PHD filtering for multi-speaker tracking," *IEEE Trans. Multimedia*, vol. 22, pp. 934–948, 2020.
[24] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[25] S. Fang *et al.*, "Attention and language ensemble for scene text recognition with convolutional sequence modeling," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 248–256.
[26] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.
[27] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4945–4949.
[28] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 316–320.
[29] Q. Kong *et al.*, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1791–1802, Nov. 2019.
[30] S. Hong, Y. Zou, W. Wang, and M. Cao, "Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 296–300.
[31] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[32] C. Wu, J. Liu, X. Wang, and X. Dong, "Object-difference attention: A simple relational attention for visual question answering," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 519–527.
[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[34] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[35] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5659–5667.
[36] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.
[37] K. Patterson, P. J. Nestor, and T. T. Rogers, "Where do you know what you know? the representation of semantic knowledge in the human brain," *Nature Rev. Neurosci.*, vol. 8, no. 12, pp. 976–987, 2007.
[38] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 1735–1742.
[39] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 776–780.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[41] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[42] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 131–135.
[43] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[44] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1003–1012.
[45] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[46] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Cheng Xue received the B.E. degree in computer science from the Henan University of Technology, Zhengzhou, China, in 2019. He is currently working toward the M.E. degree with the College of Computer Science and Electronic Engineering, Hunan University, Kaifeng, China. His research interests include audio-visual learning, computer vision, and machine learning.

Xionghu Zhong received the B.Eng. and M.Sc. degrees from Northwestern Polytechnical University, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Institute for Digital Communications, The University of Edinburgh, Edinburgh, U.K., in 2010. He was a Research Fellow with the School of Computer Engineering and a Senior Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He was with Xylem Inc., as a Data Scientist from 2017 to 2018. He is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University, Kaifeng, China. His research interests include statistical signal processing, target localization and tracking, nonparametric Bayesian modeling and machine learning methods, and their applications to distant speech enhancement and recognition, V2X communications, and water distribution network monitoring.



Minjie Cai (Member, IEEE) received the B.S. and M.S. degrees in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011, respectively, and the Ph.D. degree in information science and technology from The University of Tokyo, Tokyo, Japan, in 2016. He is currently an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University, Kaifeng, China. He is also a Cooperative Research Fellow with The University of Tokyo. His research interests include computer vision, multimedia analysis, and human-computer interaction.

Hao Chen received the B.S. degree in chemical engineering from Sichuan University, Chengdu, China, in 1998 and the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2005. He is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University, Kaifeng, China. He has authored or coauthored more than 60 articles in journals and conferences, such as the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, IPDPS, IWQoS, HiPC, and ICPP. His current research interests include parallel and distributed computing, operating systems, cloud computing, and systems security.

Wenwu Wang (Senior Member, IEEE) received the B.Sc., M.E., and Ph.D. degrees from the College of Automation, Harbin Engineering University, Harbin, China, 1997, 2000, and 2002, respectively.

He then was with King's College London, London, U.K., during 2002–2003, Cardiff University, Cardif, U.K., during 2004–2005, Tao Group Ltd. (now Antix Labs Ltd.) during 2005–2006, Creative Labs during 2006–2007, before joining the University of Surrey, Surrey, U.K., in May 2007, where he is currently a Professor of signal processing and machine learning, and a Co-Director of the Machine Audition Laboratory, Centre for Vision Speech and Signal Processing. He has coauthored more than 250 publications in his research filed, which include signal processing, machine learning and perception, with a focus on audio/speech and multimodal such as audio-visual data. He was the co-recipient of more than 10 awards, including the Judge's Award on DCASE 2020, Reproducible System Award on DCASE 2019 and DCASE 2020, and Best Student Paper Award on LVA/ICA 2018.

He is a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, and a Specialty Editor-in-Chief of the *Frontiers in Signal Processing*. He was a Publication Co-Chair for ICASSP 2019, Brighton, U.K. He is a Member of the IEEE Signal Processing Theory and Methods Technical Committee and a member of the IEEE Machine Learning for Signal Processing Technical Committee. He is also a member of the International Steering Committee of Latent Variable Analysis and Signal Separation.