Dual Attention Matching for Audio-Visual Event Localization

Yu Wu[1,2], Linchao Zhu[2], Yan Yan[3], Yi Yang[2*]
[1]Baidu Research [2]ReLER, University of Technology Sydney [3]Texas State University
yu.wu-3@student.uts.edu.au; {linchao.zhu, yi.yang}@uts.edu.au; tom_yan@txstate.edu

Abstract

In this paper, we investigate the audio-visual event localization problem. This task is to localize a visible and audible event in a video. Previous methods first divide a video into short segments, and then fuse visual and acoustic features at the segment level. The duration of these segments is usually short, making the visual and acoustic feature of each segment possibly not well aligned. Direct concatenation of the two features at the segment level can be vulnerable to a minor temporal misalignment of the two signals. We propose a Dual Attention Matching (DAM) module to cover a longer video duration for better high-level event information modeling, while the local temporal information is attained by the global cross-check mechanism. Our premise is that one should watch the whole video to understand the high-level event, while shorter segments should be checked in detail for localization. Specifically, the global feature of one modality queries the local feature in the other modality in a bi-directional way. With temporal co-occurrence encoded between auditory and visual signals, DAM can be readily applied in various audio-visual event localization tasks, e.g.*, cross-modality localization, supervised event localization. Experiments on the AVE dataset show our method outperforms the state-of-the-art by a large margin.*

1. Introduction

Multi-modal perception is essential when we human explore, capture and perceive the real world. Among these simultaneous sensory streams, vision and audio are two basic streams that convey significant information. Jointly modeling these two modalities facilitates audio-visual scenes understanding and event detection.

Recently, some works explore the cross-modal learning of visual and auditory information [3, 4, 5]. These studies focus on the representation learning of two modalities but yet to explore the temporal localization. To study how to exploit audio and visual features for event localization jointly, Tian *et al*. [30] introduce audio-visual event localization in unconstrained videos. In this problem, an audio-visual event is defined as an event that is both visible and audible in a video segment. The goal is to localize the event boundary in the temporal dimension (the *cross-modality localization* task) and to predict what category the event belongs to (the *supervised audio-visual event localization* task).

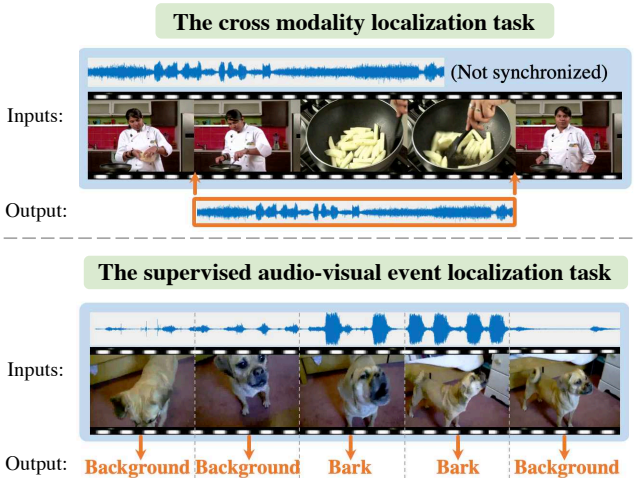

Figure 1. Examples of the audio-visual event localization problem. It includes two tasks, *i.e*., the *cross-modality localization* (CML) task and the *supervised audio-visual event localization* (SEL) task. The CML task (the upper one in the figure) is to localize the event boundary in one modality given an input event signal in the other modality. The SEL task (the lower one) is to predict the event category (including background) of each input audio-visual segment. The **orange** color in the figure indicates the output of each task.

The *cross-modality localization* (CML) task is to locate the corresponding visual signals temporally from given sound signals and vice versa. For example, as most aerial videos have no audio signals, CML is particularly useful when one needs to localize an event in aerial videos given a query audio recorded by a smartphone. As defined in [30], the task is generic, and thus no semantic label (event cate-

*This work was done when Yu Wu interned at Baidu Research. Yi Yang is the corresponding author.

gory) is provided in this task. This task aims at measuring the similarities between the two modalities, with an emphasis on generalization ability to unseen queries. In the *supervised audio-visual event localization (SEL)* task, one needs to predict which temporal segment in an input video has an audio-visual event and what category the event belongs to. We show two examples in Fig. 1.

Previous methods [30, 19] first divide a video sequence into short segments, extract visual and acoustic features for each segment. After that, they either minimize distances between segment features of the two modalities (for the CML task), or fuse the two features at the segment level (for the SEL task). The advantage of these methods is that the segment level representation reveals well about the local information, which is critical for localizing an event. The typical duration of a segment is only one second, but even a simple event may take up to a few seconds. Both visual and audio content might vary a lot within a long period. Only using local information from a small segment usually involves biases. In addition, since a segment is very short, directly fusing visual and acoustic feature at segment level is vulnerable to even a minor temporal misalignment or content noise (*e.g.*, occlusion, jitters) of the two signals. To summarize, these methods exploit the local relation between audio and vision, but overlook the *global temporal co-occurrences* between two modalities.

The global temporal co-occurrences are the correlation in a long duration between the visual and audio modalities. In an event, both visual and audio provide strong clues about the occurrence, *e.g.*, hearing a baby crying and seeing a baby in the video simultaneously. The coincidence in a long duration strongly indicates there is an event, since it is unlikely that they co-occurred across modalities merely by chance. It inspires us to take the global co-occurrences between two modalities as a strong and reliable signal while localizing an event.

We propose the Dual Attention Matching (DAM) module to leverage this relation. DAM looks into a longer video duration to better model the whole event, while also attaining local temporal information by a global cross-check mechanism. Our premise is that one must watch a longer video clip to understand the high-level event but must check shorter segments for localization. In a long duration, the audio and visual channels convey the same information about the same event, and this information should be temporally aligned. Given the global event information from one modality, DAM is designed to find which segments in the other are most relevant to the event. We model event relevance by querying from the global feature of one modality to local features in the other and vice versa.

As a module that encodes the temporal co-occurrence of audio-visual events, DAM can be readily applied in the CML and SEL task. Experiments on the AVE dataset [30] show our method outperforms the state-of-the-art methods by a large margin.

To summarize, our contributions are as follows:

- We propose Dual Attention Matching, which looks into a long duration to better model the high-level event information while also attaining local temporal information by a global cross-check mechanism.
- Our designed DAM module can be readily applied in the cross-modality localization task. Experiments show our method outperform the state-of-the-art method by a large margin.
- To address the supervised audio-visual event localization task, we design a novel joint-training framework on top of DAM. Our framework leverages both the sequence consistency of event predictions and the temporal cross-modal co-occurrence of modalities, demonstrating a decent performance in experiments.

2. Related Work

We first briefly introduce the cross modeling for vision and sound, and then discuss the applications of vision and sound techniques. Finally, we discuss related progress of our focus, the audio-visual event localization problem.

2.1. Vision and Sound Representation Learning

Recently, the cross modeling for multi modalities has attracted a lot of research attentions [3, 5, 33, 22, 23, 32, 11]. Among them, some work focus on the vision and audio classification tasks. Audio and visual information is synchronized in videos. Thus the audio channel can be used as free self-supervision. In this way, Owens *et al.* [23] leverage ambient sounds as supervision to learn visual representations. Arandjelovic and Zisserman [3] propose to learn both visual and audio representations in an unsupervised manner through an audio-visual correspondence task. In the opposite direction, Aytar *et al.* [5] propose SoundNet, which designs a visual teacher network for learning audio representations from unlabeled videos.

Based on the relation between audio and visual information, Owens and Efros [21], and Korbar *et al*. [17] concurrently propose to learn such visual and audio representation by a proxy task, the audio-visual temporal synchronization task. In the self-supervised temporal synchronization task, they train a neural network to predict whether video frames and audio are temporally aligned. We share a similar spirit with these approaches that learn from synchronized audio and visual channels in videos. Different from theirs, our primar focus is the temporal localization between two modalities. We introduce the long-term global representation to help the model to understand the event, and then check each local segment to give an accurate localization prediction.

2.2. Vision and Sound Applications

Apart from representation learning, there are also some applications of the vision and sound field.

Sound source separation. Separating the individual sound sources in an audio stream is a classic audio understanding task [7]. It is natural to introduce the visual signal to solve the problem, audio-visual source separation [21, 9, 34] . These methods enable applications ranging from playing musical instruments to speech separation and enhancement.

Audio, vision and language. Audio-visual associated between image scenes and audio captions are explored in [11]. Aytar *et al*. [6] propose to learn aligned representations across modalities, *e.g.*, audio, text, and vision. Recently, Tian *et al*. [29] propose the audio-visual video captioning task. Alamri *et al*. [1] introduce the audio-visual scene-aware dialog task, where an agents task is to answer in natural language questions about a short video.

Sound localization. The sound localization problem entails identifying which pixels or regions in a video are responsible for the recorded sound. Early works assume that a sounding object is in motion. Hershey *et al*. [13] propose to use a Gaussian process model to measure mutual information between visual motion and audio. Kidron *et al*. [16] propose to use canonical correlation analysis and exploits the spatial sparsity of audio-visual events. Recently, Senocak *et al*. [27] propose an unsupervised algorithm to address the problem of localizing the sound source in visual scenes. Arandjelovic and Zissermanl [4] locate sound source spatially in an image based on an extended correspondence network. Zhao *et al*. [34] propose PixelPlayer to separate input sounds and also locate them in the visual input.

Related to these approaches, we share the goal of locating audio in visual channels. Whereas they aim to spatially localize the audio source in videos (or images), we focus on temporally locating the audio event in the visual channel and vice versa.

2.3. Audio-visual Event Localization

Temporal event localization aims to detect and localize event in videos. Early works [12, 24] detect event in sound using only audio signals. However, the visual signals also provide rich information and should be considered in event detection. Tian *et al*. [30] propose the audio-visual event localization problem that detects events by both audio and visual modalities. In this problem, the audio-visual event may contain multiple actions or motionless sounding objects. The audio-visual event localization problem includes three tasks in [30], *i.e*., supervised and weakly-supervised audio-visual event localization, and cross-modality localization. Tian *et al*. [30] introduce an audio-guided visual attention mechanism to adaptively learn which visual regions to look for the corresponding sounding object or activity. Lin *et al*. [19] propose to integrate audio and visual feature to a global feature in a sequence-to-sequence manner. However, these methods fuse two modality features at the segment level. Differently, we propose to leverage the global event feature as the reference when localizing an event.

3. Methodology

In this section, we introduce our Dual Attention Matching (DAM) module that addresses the audio-visual event localization problem. We begin with the preliminaries of the problem statement and then introduce the DAM module in detail. In Sec 3.3 and Sec. 3.4, we illustrate how to apply our DAM module on two applications, *i.e*., the cross-modality localization (CML) task and supervised audio-visual event localization (SEL) task, respectively.

3.1. Preliminaries

In the audio-visual event localization problem, each video contains an audio-visual event that is both visible and audible. For an audio-visual video sequence $S = (S^A, S^V)$, S^A is the audio channel, and S^V is the visual channel. The temporal length of the sequence S is N seconds. Following [30], the whole video sequence is split into N non-overlapping segments $\{s_t^A, s_t^V\}_{t=1}^N$, where each segment is one second. s_t^A and s_t^V denote the audio content and the synchronized visual counterpart of the t-th segment, respectively. For a synchronized audio-visual pair (s_t^A, s_t^V), the event relevance label $y_t \in \{0, 1\}$ indicates the relevance of the two modalities about the target event. $y_t = 1$ means that the audio s_t^A and visual content s_t^V contain the event. We define the *event-relevant region* $T_E = \{t|y_t = 1, 1 \leq t \leq N\}$ as the time region when the event is happening. For each modality input, we extract the pre-trained CNN feature in the segment level. At time t, we denote f_t^A and f_t^V as the local feature (segment-level) of the audio segment and visual segment, respectively. Following [30], the local feature extractor is fixed, and we build our method on top of these local features.

3.2. Dual Attention Matching Mechanism

To obtain better event representation from one modality, we conduct a sequence embedding on the event-relevant region T_E. Given the extracted global representation of one modality, our goal is to find the local segments that are relevant to the event in the other modality, and vice versa. We use the attention mechanism to model the relation between the global feature of one modality and the local features of the other. The inner product of them is regarded as the cross-modal similarity, which is further optimized by the provided event relevance label y in training. Specifically, for candidates in the event-relevant region, we expect the local features to be close to the event representation, since they both contain information about the same event. Thus we pull the local features (of one modality) in this region

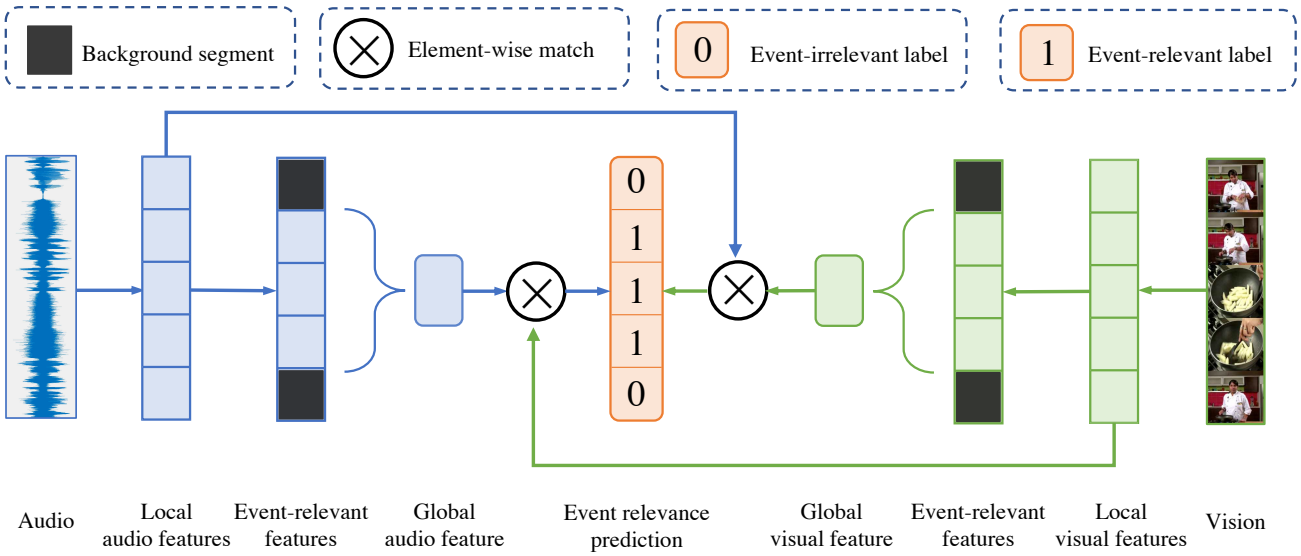

Figure 2. The proposed dual attention matching (DAM) module. DAM looks into a longer video duration to better model the high-level event information, while also attaining local temporal information by a global cross-check mechanism. DAM is optimized by finding which segments in the other are relevant to the event. We first extract the local features for each input segment and gather the features only in the event-relevant region. Then the self-attention is conducted on these local features to obtain a global event feature in this modality. To localize the event temporally, we check each local segment by calculating the dot product between the global feature (from this modality) and local feature (from the other modality). The dot product result should be 1 for those event segments and 0 for the background segments.

and the global feature (of the other) close to each other. For the rest background region, we push them away from each other. The pipeline of DAM is shown in Fig. 2.

Next, we illustrate the two components of the DAM module, *i.e.*, event-based sequence embedding, and the dual matching mechanism.

Event-based Global Feature. For an input N-length event video $\{s_t^A, s_t^V\}_{t=1}^N$, the event-relevant sequence is $S_E = \{(s_t^A, s_t^V)|t = t_1, t_2, ..., t_e\}$, where $t_i \in T_E$ indicates the index of region where an event exists, e is the length of event-relevant region T_E. To reduce the background noise, we abandon the background segments in building the global feature. Inspired by [20, 31], we apply the self-attention embedding on the event-relevant sequence to improve sequence embedding by considering the relationship among the event-relevant segments. Attention is the scaled dot-product conducted on the query, keys, and values,

$$\texttt{att}(q, k, v) = \texttt{Softmax}(\frac{qk^T}{\sqrt{d}}, v), \tag{1}$$

where d is the dimension of input feature vectors. In self-attention [31], the query q, keys k, and values v are generated by transformations of the input vector,

$$\texttt{self-att}(x) = \texttt{att}(W_q x, W_k x, W_v x), \tag{2}$$

where W_q, W_k, and W_v are the transformation weights for the input x. After the self-attention embedding, we temporally average the output features as the final representation for this modality. Taking the audio modality as an example, the event-relevant global audio representation is obtained by,

$$\phi^A(S^A) = \texttt{mean}(\texttt{self-att}(F_E^A)), \tag{3}$$

where $\texttt{mean}$ is the temporal averaged pooling operation. $F_E^A \in \mathbb{R}^{e \times d}$ denotes the concatenation of local audio features in event region T_E. In this way, we obtain the event-relevant audio representation $\phi^A(S^A) \in \mathbb{R}^d$, which contains the information about the whole event in the audio channel. Similarly, in the visual channel, we also embed the event-relevant visual feature by,

$$\phi^V(S^V) = \texttt{mean}(\texttt{self-att}(F_E^V)), \tag{4}$$

where F_E^V denotes the concatenation of local video features on region T_E. Now we have the global representations of audio and visual channels. Next, we perform the cross-modal attention matching to check each local segments.

Cross-Modal Dual Matching. The cross-modal matching is based on the assumption that information is different between event segments and background segments. In the matching, the model is trained to distinguish which segment of an auditory/visual sequence is relevant to the event. We use the dot product of global feature (in one modality)

and all the segment-level features (in the other) as the similarities (attention weights). The cross-modal matching is applied on both modalities (cross-check), *i.e.*, both from vision to audio and from audio to vision. Given the global features $\phi^A(S^A)$ and $\phi^V(S^V)$, and the local feature f_t^A and f_t^V, the event-relevant prediction is calculated by,

$$p_t^A = \sigma(\phi^V(S^V) \cdot f_t^A), \tag{5}$$

$$p_t^V = \sigma(\phi^A(S^A) \cdot f_t^V), \tag{6}$$

where p_t^A and p_t^V denote the event relevance predictions on the audio and visual channel at the t-th segment, respectively. σ is the Sigmoid activation function that converts the dot product to the range $(0, 1)$. With these two cross-modal matching, we have the final event-relevant prediction by,

$$p_t = \frac{1}{2}(p_t^A + p_t^V). \tag{7}$$

The ground truth for the event-relevant prediction task is the event-relevant label y_t, *i.e.*, p_t should be 1 for if segment t is in the event-relevant region T_E, and 0 for the background region. We use the Binary Cross Entropy (BCE) loss to optimize the DAM module.

On top of the DAM module, we design frameworks for two audio-visual event localization applications, *i.e.*, the CML task, and the SEL task.

3.3. Cross-Modality Localization

In the cross-modality localization (CML) task, given a few event-relevant segments of one modality, the target is to find the position of its synchronized content in the other modality. This task is suitable to evaluate the model's ability of leveraging audio-visual connections, since correlations are the only information that can be used to localize the event in the target modality.

The CML task contains two directional localization, *i.e.*, visual localization from audio (A2V) and audio localization from visual content (V2A). In the A2V task, given a l-second event-relevant audio sequence $\hat{S}^A$ from $\{s_t^A\}_{t=1}^N$, where $l < N$, the target to find its synchronized l-second visual segment within $\{s_t^V\}_{t=1}^N$. As defined in [30], there is no semantic label (event category) provided during localization. Similarly, in the V2A task, given a l-second visual segment $\hat{S}^V$, we would like to find its l-second audio segment within $\{s_t^A\}_{t=1}^N$.

Our designed DAM module can be readily applied to the CML task. In training, the whole video and event relevance labels are provided. We then train the DAM module as discussed in Sec. 3.2. In the inference stage, we first obtain the event-based global feature from the query sequence, and then use the global feature as a query to check each local segment of the candidate. Each segment is assigned a prediction score that indicates its relevance with the input query. Finally, we look into the prediction scores of the N-length candidate segments, and output the l-length sequence with maximum contiguous sum as the final localization prediction.

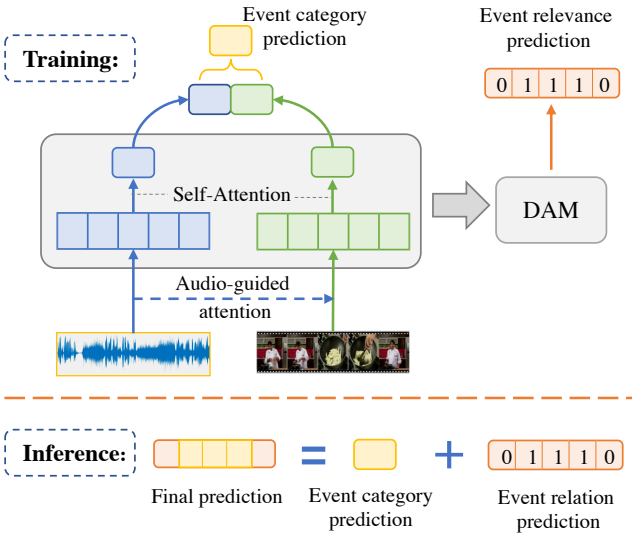

Figure 3. Framework for the *Supervised Audio-Visual Event Localization* task. The yellow block and orange block indicate the event *category* prediction and event *relevance* prediction, respectively. For inference, the final is the combination of the two prediction.

3.4. Supervised Audio-Visual Event Localization

The supervised audio-visual event localization task is to predict which temporal segment of an input video has an audio-visual event and *what category the event belongs to*. In this task, we have both the event-relevant region annotations y and the event category label annotations y^c. Note that only one event category exists within a video in the task. The target is to predict the categories (including background) for all the N-length segments of an input event video.

Different from [30, 19], We decouple this task by two subtasks, *i.e.*, 1) predicting the event category based on the overall sequences, and 2) differentiate background segments in the untrimmed event videos. As shown in Fig. 3, the model mainly contains two branches. We extract the global representation of the audio channel and visual channel by the self-attention mechanism. Note that the self-attention takes as input all segments including background. The reason is that we cannot access the annotation of event region T_E during evaluation. Then we fuse the two global features and predict the *event category* $\hat{y}^c$ based on the fused features. In the meantime, the DAM module takes the global features and check each local (segment-level) features to predict *event relevance* $\hat{y}_t$, which is further used to determine whether the t-th segment is the background. Fol-

lowing [30], we also use audio-guided visual attention in generating local visual features. In the inference stage, the final prediction is the combination of predictions $\hat{y}^c$ and $\hat{y}_t$. For a t-th segment, if $\hat{y}_t < 0.5$, the final prediction for this segment is background. If $\hat{y}_t \geq 0.5$, the segment is predicted to be event-relevant and thus the final prediction is the event category prediction $\hat{y}^c$.

In the training stage, we have the corresponding event category label and event relevance label, thus the overall objective function is,

$$\mathcal{L} = \lambda \mathcal{L}^c + (1-\lambda)\frac{1}{N}\sum_{t=1}^{N}\mathcal{L}_t^r, \tag{8}$$

where $\mathcal{L}^c$ is the Cross-Entropy loss for the event category prediction $\hat{y}^c$, and $\mathcal{L}_t^r$ is the Binary Cross Entropy loss for the event relevance prediction $\hat{y}_t^r$ at t-th segment. We will evaluate the effectiveness of λ in Sec. 4.3.

4. Experiments

We first discuss the experimental setups and then compare our method with the state-of-the-art methods on the AVE dataset under two tasks. Ablation studies and qualitative results are provided to show the effectiveness of DAM.

4.1. Experiment Setup

The Audio-Visual Event (AVE) dataset [30] derived from AudioSet [10], contains 4,143 videos covering 28 event categories. The videos in the AVE dataset involve a wide range of audio-visual event domains, *e.g.*, human activities, animal activities, music performances, and vehicle sounds. The detailed events categories, including man speaking, dog barking, playing guitar, and frying food *etc.*, last at least two-second in length for each video. Each video lasts 10 seconds with both audio and video tracks. Videos in AVE are temporally labeled with audio-visual event boundaries, which demonstrates whether a segment is event-relevant or the background.

Evaluation metrics. In the CML task, the only information provided in training is the audio-visual event boundaries. The task has two evaluation subtasks, including visual localization from audio (A2V) and audio localization from visual content (V2A). A good matching defined in this task is that a matched audio/visual segment is exactly the same as its ground truth; otherwise, it will be a bad matching. We compute the percentage of good matchings overall all testing samples as prediction accuracy to evaluate the performance of CML. In the SEL task, we predict the category for each one-second segment in an input video. Note "background" is also a category in this classification task. The overall classification accuracy is used as an evaluation metric for this task.

Method	A2V	V2A	Average
DCCA [2]	34.1	34.8	34.5
AVDLN [30]	35.6	44.8	40.2
Ours	**47.1 ± 1.6**	**48.5 ± 1.4**	**47.8 ± 1.5**

Table 1. Comparisons with the state-of-the-art methods on the cross-modality localization task. A2V: visual localization from audio sequence query; V2A: audio localization from visual sequence query. "Average" indicates the averaged score of two tasks. We report the mean and standard deviation of three runs to reduce randomness.

Implementation details. We adopt pre-trained CNN models to extract local segments features for audio and visual content. For a fair comparison, we use the VGG-19 [28] network pre-trained from the ImageNet [25] dataset as the visual CNN model to extract features for each 1-second visual segment. Similarly, for audio representation, we extract the audio representation for each 1 second audio segment via a VGG-like network [14] pre-trained on AudioSet [10]. In experiments, for a fair comparison, we use the same low-level structure (*e.g.*, low-level embedding, segment-level attentions) as used in [30]. For the self-attention module, we use the default structure as illustrated in [31].

4.2. Comparison with State-of-the-art Results

Cross-modality localization. Table 1 shows the performances of our method and state-of-the-art methods AVLN [30] and DCCA [2] on the CML task. The AVLN method, similar with [4], extract features for two modalities and measures the relativeness of them by a simple Euclidean distance. Different from AVLN [30] and DCCA [2] that only focus on the local segments, our DAM first watches a long event sequence to obtain a stable representation and then check each segment for a better localization. Our method, with the designed DAM module, outperforms the state-of-the-art methods by a large margin on both A2V and V2A tasks. Specifically, on the A2V task, our method improves the accuracy from 35.6% to 47.1%. The improvement is not easy since the CML task is challenging. The provided annotations in this task are very limited (only with the event boundaries but without the event label), the contents are very different (one is audio, and the other is visual), and the evaluation metric is strict (which counts only the exact matches).

Supervised audio-visual event localization. We also test our proposed framework on the SEL task, which is a segment-level event classification problem. In training, we have detailed event categories (including background) annotations for each 1-second segment. We compare our method with state-of-the-art methods. ED-TCN [18] is a state-of-the-art temporal action labeling method. Tian *et al*. [30] propose the baseline for this task, which utilizes pre-trained CNN models to encode audio and visual inputs,

Method	Accuracy (%)
ED-TCN [18]	46.9
Audio (pre-trained VGG-like [14])	59.5
Visual (pre-trained VGG-19 [28])	55.3
Audio-visual [30]	71.4
AVSDN* [19]	72.6
Audio-visual+Att [30]	72.7
Ours	**74.5 ± 0.6**

Table 2. Comparisons with the state-of-the-art methods in the supervised audio-visual event localization task on the AVE dataset. * indicates the reproduced performance using the same pre-trained VGG-19 feature for a fair comparison. We report the mean and standard deviation of three runs to reduce randomness.

Method	V2A	A2V	Average
DAM w/ RNN	41.8	47.9	44.9
DAM w/ Averaged Pooling	46.0	46.1	46.1
DAM w/ Max Pooling	45.8	46.2	46.0
DAM w/ LSTM [15]	43.5	48.1	45.8
DAM w/ GRU [8]	45.5	47.4	46.5
DAM w/ BLSTM [26]	44.2	48.1	46.2
DAM w/ Self-Att [31]	47.1	48.5	47.8

Table 3. Comparisons of different sequence embedding functions used in the DAM module on the cross-modality localization task.

adapts LSTM to capture temporal dependencies, and applies a fully connected layer to make the final predictions. On top of the baseline model, Tian *et al*. [30] further introduce the audio-guided visual attention mechanism to adaptively learn which visual regions to look for the corresponding sounding object or activity. Lin *et al*. [19] propose the AVSDN method by introducing an additional LSTM to replace the final prediction classifier. Table 2 summarizes the performances of our method and the state-of-the-art methods on the AVE dataset. We observe that our method yields higher accuracy than the best state-of-the-art result (74.5% versus 72.7%).

4.3. Ablation Studies

Different sequence embedding functions. We systematically investigate different sequence embedding functions to replace the self-attention module (Eqn. (2)) used in DAM. The common sequence embedding functions that model the sequence relationship are Averaged Pooling, Max Pooling, RNN, LSTM [15], Bidirectional-LSTM [26], GRU [8], and Self-Attention [31]. We evaluate these sequence embedding functions in our DAM module to reveal the effect of global information. The performance comparisons are reported in Table 3. Among all the embedding functions, self-attention achieves the best performance. It is worth mentioning that with two non-parametric embedding functions (Max Pooing and Averaged Pooling), the overall performance of the DAM module still outperforms the state-of-the-art method [30] that only focuses on local segments. It is consistent with our motivation that one must watch a long video clip to understand the whole event before checking each local segment for localization.

Method	V2A	A2V	Average
DAM w/ Self-Matching	28.6	29.8	29.2
DAM w/ Cross-Matching	47.1	48.5	47.8

Table 4. Comparisons of the cross-modal matching and self-matching on the cross-modality localization task. "Self-Matching" indicates we use the global feature of the modality itself as a query to match the local features. "Cross-Matching" is the cross-modal matching in our DAM (discussed in Sec. 3.2).

Method	Accuracy (%)
Ours w/o Matching	70.7
Ours w/ Self-Matching	74.2
Ours w/ Cross-Matching	74.5

Table 5. Ablation studies on the matching mechanism on the supervised audio-visual event localization task. "Ours w/o Matching" indicates our framework without the DAM module. "Self-Matching" indicates we use the global feature of the modality itself as a query to match the local features.

Cross-modal matching versus self-matching. We also perform self-matching instead of cross-modal matching in the DAM to validate the effectiveness of cross-checking. Specifically, instead of using the global feature from the other modality, we change the global feature $\phi^V(S^V)$ and $\phi^A(S^A)$ in Eqn. 5 and Eqn. 6 by the global feature of the modality itself. Table 4 reports the performance comparison in the CML task. "Ours w/ Self-Matching" indicates the model is learned by leveraging the weak relation within the modality itself. In the inference stage, we calculate the Cosine distance between the query and candidates, and output the one with the minimum distance as the localization prediction. The performance of self-matching is far away from our DAM, indicating that the temporal co-occurrence is a strong correlation between modalities in CML. Table 5 summarise the performance comparisons on the SEL task. The "Ours w/o Matching" denotes the framework that utilizes the consistency of event sequence, but do not use global feature to check local segments via DAM. Since the model cannot distinguish the background and event segments, it thus achieves a poor performance (70.7%) in the SEL task. The self-matching model outperforms the one without matching by 3.5 points, which also validates our motivation that it is helpful to watch the whole event before localizing each small segments. Our DAM with cross-matching further improves the performance by leveraging cross modal information. Note the performance gain in the SEL task is relatively small compared to that in the CML

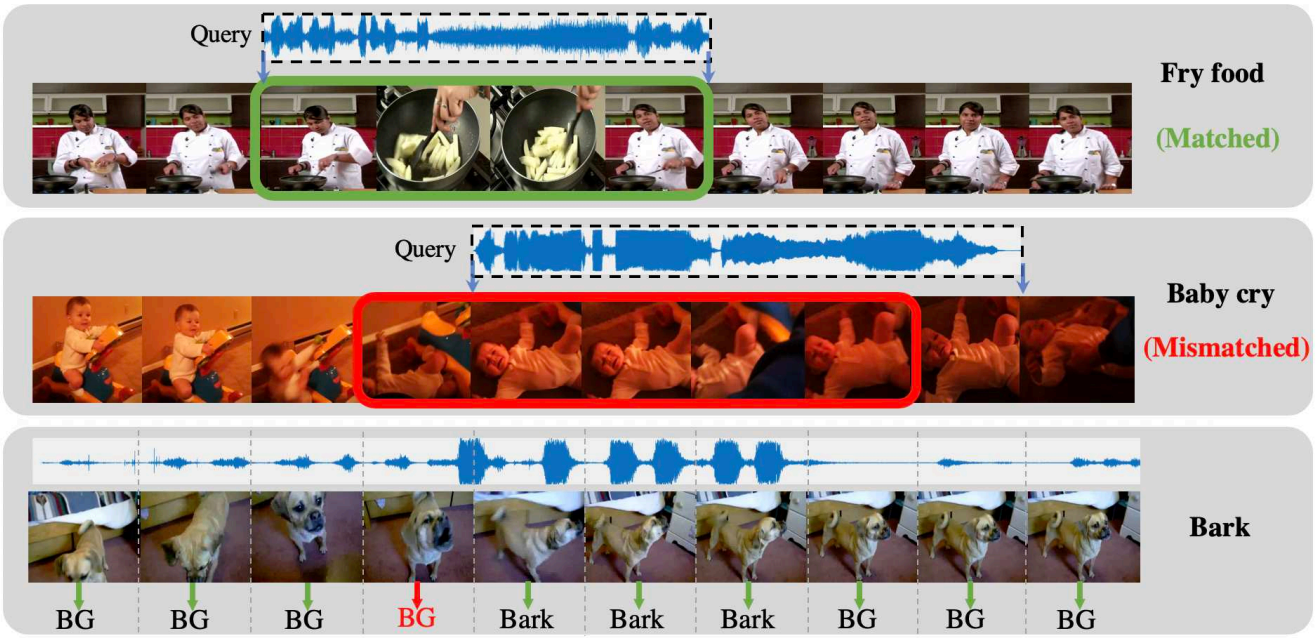

Figure 4. Qualitative results. Green and red stand for the correct and wrong predictions, respectively. The first two examples show the A2V task, in which the input audio query is located in its ground truth position on the temporal dimension (horizontal). The bottom one shows an example of the SEL task. The model fails in predicting the fourth segment. The ground truth for the fourth segment is "barking" while our model predicts it to be "BG" (background).

task. The reason is that the complementarity of the two modalities has already been utilized through the fused features of the two modalities during prediction.

Analysis on the balancing parameter λ. In Eqn. (8), λ is a hyper-parameter balancing the contributions of the event relevance loss $\mathcal{L}^r$ and the event category loss $\mathcal{L}^c$. Fig. 5 shows the performance curves over different values of λ. Note $\lambda = 1$ denotes only using the event category loss during the training, *i.e.*, without the DAM module. The best performance is achieved at $\lambda = 0.5$ with 74.5%.

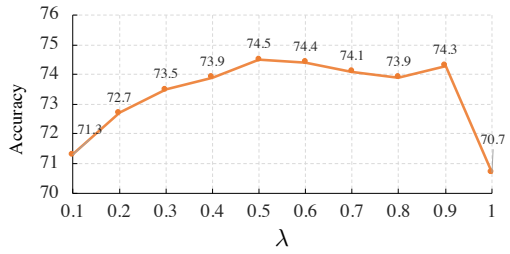

Figure 5. Performances on different values of the balancing parameter λ (defined in Eqn. (8)) on the SEL task.

4.4. Qualitative Results

We show some qualitative results of our DAM model in Fig. 4. Green and red in this figure stand for the correct and wrong predictions, respectively. The first two examples show the A2V task. We draw the input audio query in its ground truth position on the temporal dimension (horizontal). For the first example, although the visual content varies a lot, our DAM still succeed in finding the correct temporal location given a sound about frying food. The second example is more hard. Given a sound of baby crying, it is not easy to localize its visual segments because facial motion is too small. Therefore, the predicted results (red box) is mismatched with the query. The bottom one shows an example of the SEL task. The model fails in predicting the category for the fourth segment. The ground truth is "barking" while our model predicts it to be "background". The main reason is that "barking" starts from the middle of this segment.

5. Conclusion

In this work, we investigate the audio-visual event localization problem and propose the dual attention matching (DAM) module. Different from previous methods that focus on local segments, our DAM looks into a longer video duration to better model the high-level event information while also attaining local temporal information by a global cross-check mechanism. Our intuition is that one must watch a longer video clip to understand the high-level event but must check shorter segments for localization. Specifically, given the global event information from one modality, DAM is designed to find which segments in the other are most relevant to the event. We model event relevance through querying from the global feature to the local feature. Experiments show our method outperforms the state-of-the-art methods by a large margin.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio-visual scene-aware dialog. In *CVPR*, 2019.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

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

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

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for single-microphone speaker separation. In *ICASSP*, 2017.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

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

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[11] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *NIPS*, 2016.

[12] Toni Heittola, Annamaria Mesaros, Antti J. Eronen, and Tuomas Virtanen. Context-dependent sound event detection. *EURASIP J. Audio, Speech and Music Processing*, 2013.

[13] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 2000.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[17] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NIPS*, 2018.

[18] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017.

[19] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*, 2019.

[20] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.

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

[22] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016.

[23] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016.

[24] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *ICASSP*, 2016.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[26] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

[27] 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.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014.

[29] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872*, 2018.

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

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[32] Yu Wu, Lu Jiang, and Yi Yang. Revisiting embodiedqa: A simple baseline and beyond. *arXiv preprint arXiv:1904.04166*, 2019.

[33] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM MM*, 2018.

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