Cross-Modal Relation-Aware Networks for Audio-Visual Event Localization

Haoming Xu*†
South China University of Technology
Guangzhou, China
xuhaoming.cs@gmail.com

Runhao Zeng*
South China University of Technology
Guangzhou, China
runhaozeng.cs@gmail.com

Qingyao Wu*
South China University of Technology
Guangzhou, China
qyw@scut.edu.cn

Mingkui Tan‡
South China University of Technology
Guangzhou, China
mingkuitan@scut.edu.cn

Chuang Gan
MIT-IBM Watson AI Lab
Cambridge, USA
ganchuang1990@gmail.com

ABSTRACT

We address the challenging task of event localization, which requires the machine to localize an event and recognize its category in unconstrained videos. Most existing methods leverage only the visual information of a video while neglecting its audio information, which, however, can be very helpful and important for event localization. For example, humans often recognize an event by reasoning with the visual and audio content simultaneously. Moreover, the audio information can guide the model to pay more attention on the informative regions of visual scenes, which can help to reduce the interference brought by the background. Motivated by these, in this paper, we propose a relation-aware network to leverage both audio and visual information for accurate event localization. Specifically, to reduce the interference brought by the background, we propose an audio-guided spatial-channel attention module to guide the model to focus on event-relevant visual regions. Besides, we propose to build connections between visual and audio modalities with a relation-aware module. In particular, we learn the representations of video and/or audio segments by aggregating information from the other modality according to the cross-modal relations. Last, relying on the relation-aware representations, we conduct event localization by predicting the event relevant score and classification score. Extensive experimental results demonstrate that our method significantly outperforms the state-of-the-arts in both supervised and weakly-supervised AVE settings. The source code is available at https://github.com/FloretCat/CMRAN.

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding*.

KEYWORDS

Audio-visual event localization; Cross-modality relation; Attention

ACM Reference Format:
Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. 2020. Cross-Modal Relation-Aware Networks for Audio-Visual Event Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413581

1 INTRODUCTION

Event localization is an important yet challenging task for video understanding, which requires the machine to localize events/actions and recognize the categories in an unconstrained video. In recent years, this task has attracted increasing attention [3, 10, 31, 39, 40, 42]. Most existing methods take only RGB frames or optical flow as input to localize and identify an event. However, due to the strong visual background interference and large visual content variations, it can be difficult to localize events with only visual information. Hearing is one of the main ways for humans to perceive the real world, and the audio signal is potentially helpful for event localization. Specifically, audio signals can guide us to focus on informative regions of visual scenes. This is intuitively apparent since when we hear the sounds of a dog barking in a video, we are most likely to observe the area where the dog is located. Besides, the audio signals often carry useful cues for reasoning.

The audio-visual event (AVE) localization task, which requires a machine to determine the presence of an event that is both audible and visible in a video segment and to what category the event belongs, has attracted increasing attention. In this paper, we study how to effectively leverage audio and visual information for event localization by addressing the AVE localization task. However, this task is very challenging due to the following difficulties: 1) complex visual backgrounds in an unconstrained video make it difficult to localize an AVE, and 2) localizing and recognizing an AVE requires the machine to simultaneously consider the information from two modalities (*i.e.*, audio and vision) and exploit their relations. This is

*Authors contributed equally.
†Also with PengCheng Laboratory, Shenzhen, China
‡Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413581

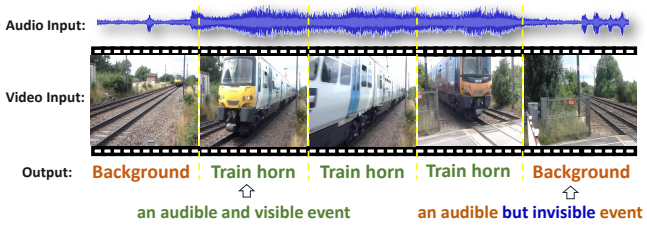

Figure 1: An illustrative example of the audio-visual event localization. Given a video with acoustic contents, we seek to determine whether there exists an event that is both audible (*e.g.*, hearing a sound emitted by a moving train) and visible (*e.g.*, seeing the moving train) in any segment. To accomplish this task, we need to consider the two modalities and capture the relations between them, which is very challenging.

difficult since it is nontrivial to build connections between complex visual scenes and intricate sounds [25]. Existing methods [24, 30, 36] in this task process two modalities independently, and simply fused them together just before the final classifiers. They focus on modeling internal temporal relations (*i.e.*, intra-modality relations) to exploit potential cues for event localization, while neglecting rich and valuable inter/cross-modality relations.

We contend that the cross-modality relation also plays an important role in AVE localization. Intuitively, the cross-modality relation is the audio-visual correlation between audio and video segments. For example, as shown in Figure 1, we hear the sound of a train horn while seeing a moving train. This audio-visual correlation suggests an event that is audible and visible. Therefore, we argue that cross/inter-modality relations also contribute to the detection of an audio-visual event. Such intuition and case motivate us to exploit cross/inter-modality relations for AVE localization.

In recent years, self-attention mechanism [32] has achieved great success in capturing intra-modality relations among words in NLP [5, 6]. It first transforms input features into query, key and value (*i.e.*, memory) features. Then, it calculates the attentive output using a weighted summation over all values in the memory, where the weights (*i.e.*, relations) are learned from the key in the memory and the query. However, since the query and memory are derived from the same modality, directly applying self-attention to event localization cannot exploit the cross-modality relations between visual and acoustic contents. On the contrary, if the memory acquires features of two modalities, then the query (from one of the two modalities) will enable exploration of the cross-modality relations while not missing the intra-modality relation information.

In this paper, we propose a relation-aware module to build connections between visual and audio information by exploiting inter-modality relations. This module wraps an attention mechanism called cross-modality relation attention (CMRA) inspired by the success of self-attention. Different from self-attention, the query is derived from one modality while the keys and values are derived from two modalities in our CMRA. In this way, an individual segment from a modality can aggregate useful information from all related segments from two modalities based on the learned intra- and inter-modality relations. Our intuition is that simultaneously watching the visual scenes and listening to the sounds (*i.e.*, exploiting intra- and inter-modality relation information from two modalities) is more effective and efficient than separately perceiving them (*i.e.*, exploiting only intra- or inter-modality relation information) for localizing an audible and visible event. We attempt to exploit both useful relations to facilitate representation learning and further boost the performance of AVE localization.

Besides, since the strong visual background interference can obstruct accurate event localization, we aim to highlight informative visual regions and features to reduce the interference. To this end, we propose an audio-guided spatial-channel attention module, which leverages audio information to build visual attention at spatial and channel levels. We integrate these components together and present a cross-modal relation-aware network, which outperforms state-of-the-arts by a certain margin in supervised and weakly-supervised AVE localization tasks on AVE dataset.

Our main contributions in this paper are as follows:

- We propose an Audio-Guided Spatial-Channel Attention module (AGSCA) to leverage the guidance capability of audio signals for visual attention, which precisely highlights informative features and sounding regions.
- We propose a relation-aware module to exploit the intra-modality and inter-modality relations for event localization.
- Built upon these two modules, we present a Cross-Modal Relation-Aware Network for supervised and weakly supervised AVE localization tasks. Experimental results demonstrate that our method significantly outperforms the state-of-the-arts in both tasks on AVE dataset, suggesting the effectiveness of the proposed method.

2 RELATED WORK

In this section, we first introduce recent works on audio-visual learning. Then, we narrow the scope to the task we investigate in this paper *i.e.*, audio-visual event localization. Last, we briefly introduce the attention mechanism.

Audio-Visual Learning. Audio-visual learning has attracted attention from many domains, such as action recognition [12, 20, 27], sound source separation [9, 43, 44] and generation [8], and audio-visual event localization [24, 30, 36]. Among them, Gao *et al.* [12] use audio to build a previewing mechanism to reduce temporal redundancies. Kazakos *et al.* [20] propose a sparse temporal sampling strategy to fuse multiple modalities to boost action recognition. Owens *et al.* [27] propose to use audio as a supervisory signal for learning visual models in an unsupervised manner. Oh *et al.* [26] present a Speech2Face framework that uses the voice-face correlations to generate facial images behind the voice. In addition, to exploit the readily available large-scale unlabelled videos, many works [1, 11, 28] leverage audio-visual correspondence to learn audio-visual representations in a self-supervised manner.

Audio-Visual Event Localization. Tian *et al.* [30] first use two LSTMs to separately model temporal dependencies of audio and video segment sequences and then fuse audio and visual features via additive fusion for event category prediction. Lin *et al.* [24] first separately process audio and visual modalities and then fuse features of two modalities via LSTMs, which works in a sequence-to-sequence manner. Wu *et al.* [36] propose a dual attention matching module, which uses global information obtained by intra-modality relation

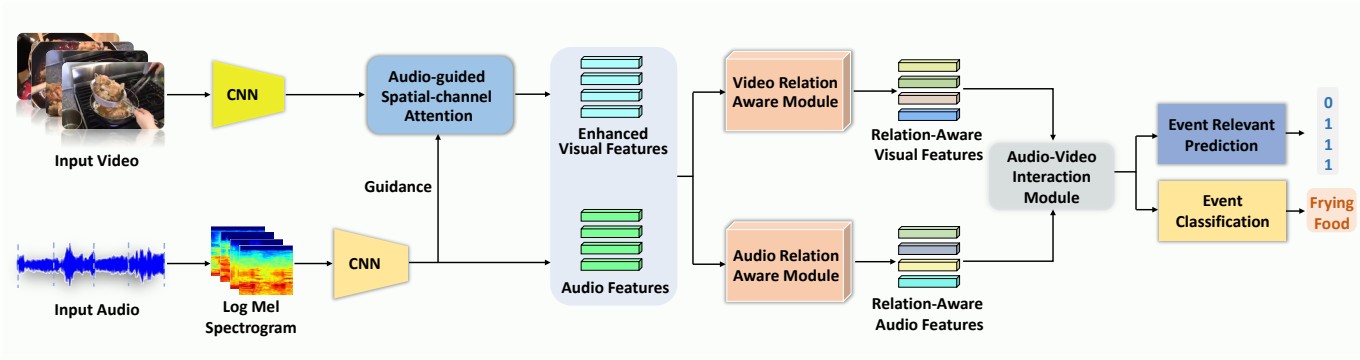

Figure 2: Schematic of our cross-modal relation-aware network. First, audio-guided spatial-channel attention serves to leverage audio information to guide visual attention at spatial and channel levels. Then, two relation-aware modules capture both intra-modality relations and inter-modality relations for two modalities separately. Last, cross-modal relation-aware visual and acoustic features are incorporated together via an audio-video interaction module, yielding a joint dual-modality representation for the following classifiers.

modeling and local information to measure cross-modality similarity via the inner-product operation. The cross-modality similarity directly serves as a final event relevance prediction.

These methods mainly concentrate on leveraging intra-modality relations as potential cues, ignoring the equally valuable cross-modality relation information for event localization. Different from these methods, our proposed cross-modal relation-aware networks enable bridging connections between visual and audio modalities, by simultaneously exploiting both the intra- and inter-modality relation information.

Attention Mechanism. Attention mechanism mimics human visual perception function. It seeks to automatically focus on certain portions of the input that have high activation. Attention mechanism has many variants, and our work mainly relates to self-attention.

Vaswani *et al.* [32] propose the self-attention mechanism to capture long-range dependencies among words for machine translation. Recently self-attention mechanism has emerged widely in NLP [5–7, 35] and many language-vision tasks [4, 18, 19, 38, 41]. Devlin *et al.* [6] propose the well-known BERT model based on self-attention for pretraining word embeddings. Wang *et al.* [33] transfer self-attention into the vision domain, and attempt to capture pixel-level long-range dependencies spatially and temporally. Ye *et al.* [37] propose the cross-modal self-attention mechanism for referring image segmentation, where they perform self-attention over the fused features from multiple modalities. Hu *et al.* [16] propose a relation module based on self-attention to capture the relations among bounding boxes for object detection.

In contrast to the self-attention, which focuses on capturing relations within a modality, our proposed cross-modality relation attention enables simultaneous exploitation of intra- and inter-modality relations for audio-visual representation learning.

3 PROPOSED METHOD

Notations. Let $\mathcal{S} = \{S_t = (V_t, A_t)\}_{t=1}^{T}$ be a video sequence with T non-overlapping segments. Here, V_t and A_t represent the visual content and its corresponding audio content of the t-th segment, respectively.

Problem Definition. Given a video sequence $\mathcal{S}$, AVE localization requires a machine to predict the event label (including background) for each segment S_t relaying on V_t and A_t, as illustrated in Figure 1. An audio-visual event is defined as an event that is both audible and visible (*i.e.*, hearing a sound emitted by an object and simultaneously seeing the object). If a segment S_t is not both audible and visible, it should be predicted as background, as labeled in Figure 1. The main challenge in this task is that the machine is required to analyze two modalities and capture their relations. Cross-modality relation information can be used to boost performance and, unfortunately, is mostly ignored by existing methods. In this paper, we study this task in two settings:

Supervised Setting. We have access to segment-level labels during the training phase. A segment-level label indicates the category (including background) of the corresponding segment. Non-background category labels are given only if the sounds and the corresponding sounding objects are presented.

Weakly supervised Setting. We can access only video-level labels during training, and we still aim to predict a category for each segment during testing. A video-level label indicates whether a video contains an audio-visual event and to what category the event belongs.

3.1 General Scheme

We focus on solving the problem that most existing event localization methods neglect the information from the audio signal in a video, which, however, can help to alleviate the interference of complex background and provide more cues for reasoning. We study a method that leverages both the visual and audio information for event localization and evaluate it on an audio-visual event localization task [30], which requires the machine to localize an event that is both audible and visible in an untrimmed video. This task is very challenging, since an unconstrained video often contains complex backgrounds and it is nontrivial to build connections between complex visual scenes and intricate sounds [25]. To address

these challenges, we propose an audio-guided attention module to highlight certain spatial regions and features to reduce background interference. We also devise relation-aware modules to exploit inter-modality relations along with intra-modality relations for localizing an audio-visual event.

Specifically, we propose a cross-modal relation-aware network (CMRAN) with three main components as shown in Figure 2, including audio-guided spatial-channel attention module (AGSCA), relation-aware module and audio-video interaction module. Given a video sequence $\mathcal{S}$, we first forward each audio-visual pair $\{V_t, A_t\}$ through pretrained CNN backbones to extract segment-level features $\{\boldsymbol{v}_t, \boldsymbol{a}_t\}_{t=1}^{T}$. Then, we forward audio and visual features through the AGSCA module, to obtain enhanced visual features. With audio features and enhanced visual features at hand, we prepare two relation-aware modules (*i.e.*, video relation-aware module and audio relation-aware module in Figure 2) separately for audio and visual features. We feed visual and audio features into the relation-aware modules to exploit both relations for two modalities. Last, the relation-aware visual and audio features are fed into the audio-video interaction module, yielding a comprehensive joint dual-modality representation for event classifiers.

In the following, we first illustrate the audio-guided spatial-channel attention in Section 3.2. We then introduce the relation-aware module and audio-video interaction module in Section 3.3 and Section 3.4, respectively. Last, we introduce how to apply our method to both supervised and weakly-supervised AVE localization in Section 3.5.

3.2 Audio-Guided Spatial-Channel Attention

Previous works [27, 30] have shown that audio signals are capable of guiding visual modeling. Besides, the channel attention can ignore irrelevant features and improve the quality of visual representations [17]. Inspired by these, we propose an audio-guided spatial-channel attention module (AGSCA), which seeks to make the best of the audio guidance capability for visual modeling. Different from [30], where audio features only participate in visual attention in the spatial dimension, our AGSCA exploits audio signals to guide visual attention in both spatial and channel dimensions, which emphasizes informative features and spatial regions to boost the localization accuracy. We follow [2] to perform channel and spatial attention sequentially.

Given audio features $\boldsymbol{a}_t \in \mathbb{R}^{d_a}$ and visual features $\boldsymbol{v}_t \in \mathbb{R}^{d_v \times (H*W)}$ where H and W are the height and width of feature maps respectively, AGSCA first generates channel-wise attention maps $\boldsymbol{M}_t^c \in \mathbb{R}^{d_v \times 1}$ to adaptively emphasize informative features. It then produces spatial attention maps $\boldsymbol{M}_t^s \in \mathbb{R}^{1 \times (H*W)}$ for the channel-attentive features to highlight sounding regions, yielding channel-spatial attentive visual features $\boldsymbol{v}_t^{cs}$, as illustrated in Figure 3. The attention process can be summarized as,

$$\begin{aligned} \boldsymbol{v}_t^{cs} &= \boldsymbol{M}_t^s \otimes (\boldsymbol{v}_t^c)^T , \\ \boldsymbol{v}_t^c &= \boldsymbol{M}_t^c \odot \boldsymbol{v}_t , \end{aligned} \tag{1}$$

where $\otimes$ denotes matrix multiplication, and $\odot$ means element-wise multiplication. We next separately introduce the channel-wise attention that generates attention maps $\boldsymbol{M}_t^c$ and spatial attention that produces attention maps $\boldsymbol{M}_t^s$.

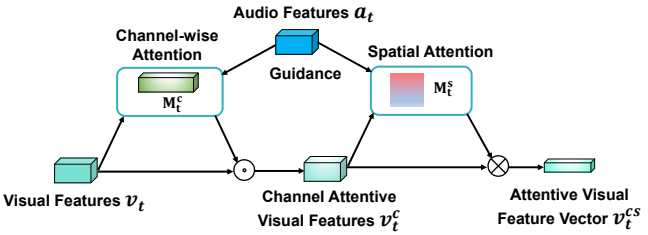

Figure 3: Illustration of the AGSCA module. AGSCA leverages the audio guidance capability to guide visual attention at channel level (left part) and spatial level (right part).

Channel-Wise Attention. We explicitly model the dependencies between channels of features with the guidance of audio signals. Specifically, we first project audio and visual features to the same dimension d_v using fully-connected layers with non-linearity, resulting in audio guidance maps $\boldsymbol{a}_t^m \in \mathbb{R}^{d_v}$ and projected visual features with dimensions of $d_v \times (H * W)$. We then leverage the guidance information of $\boldsymbol{a}_t^m$ by fusing visual features with $\boldsymbol{a}_t^m$ via element-wise multiplication. Following [17], we spatially squeeze the fused features by global average pooling. Last, we forward the fused feature vector through two fully-connected layers with non-linearity to model the relationships between channels, yielding channel attention maps $\boldsymbol{M}_t^c$. We give the details as follows:

$$\boldsymbol{M}_t^c = \sigma(\boldsymbol{W}_1 \boldsymbol{U}_1^c (\delta_a(\boldsymbol{U}_a^c \boldsymbol{a}_t \odot \boldsymbol{U}_v^c \boldsymbol{v}_t))) , \tag{2}$$

where $\boldsymbol{U}_a^c \in \mathbb{R}^{d_v \times d_a}$, $\boldsymbol{U}_v^c \in \mathbb{R}^{d_v \times d_v}$, and $\boldsymbol{U}_1^c \in \mathbb{R}^{d \times d_v}$ are fully-connected layers with ReLU as an activation function, $\boldsymbol{W}_1 \in \mathbb{R}^{d_v \times d}$ are learnable parameters with $d = 256$ as a hidden dimension, δ_a indicates global average pooling, and σ denotes the sigmoid function. We add a residual connection by adding one to each element of $\boldsymbol{M}_t^c$ to obtain the final channel attention maps.

Spatial Attention. We also leverage the guidance capability of audio signals to guide visual spatial attention. Spatial attention follows a similar pattern to the aforementioned channel-wise attention. Note that the input visual features $\boldsymbol{v}_t^c$ are channel attentive. We formulate the process of spatial attention as follows:

$$\begin{aligned} \boldsymbol{M}_t^s &= \text{Softmax}(\boldsymbol{x}_t^s) , \\ \boldsymbol{x}_t^s &= \delta(\boldsymbol{W}_2((\boldsymbol{U}_a^s \boldsymbol{a}_t) \odot (\boldsymbol{U}_v^s \boldsymbol{v}_t^c))) , \end{aligned} \tag{3}$$

where $\boldsymbol{U}_a^s \in \mathbb{R}^{d \times d_a}$, $\boldsymbol{U}_v^s \in \mathbb{R}^{d \times d_v}$ are fully-connected layers with ReLU as an activation function, $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times d}$ are learnable parameters with $d = 256$ as a hidden dimension, and δ denotes the hyperbolic tangent function. With the spatial attention maps $\boldsymbol{M}_t^s$ at hand, we perform weighted summation over $\boldsymbol{v}_t^c$ according to $\boldsymbol{M}_t^s$ to highlight informative regions and shrink spatial dimensions, yielding a channel-spatial attentive visual feature vector $\boldsymbol{v}_t^{cs} \in \mathbb{R}^{d_v}$ as output.

3.3 Relation-Aware Module

A relation-aware module involves a cross-modality relation module denoted as M_{cmra}, and an internal temporal relation module denoted as M_{self}. The module M_{cmra} contains the cross-modality relation attention mechanism (CMRA) illustrated below to exploit relation information. M_{self} serves as an assistant of M_{cmra}. We first give an overall description of the relation-aware module and then detail each component separately.

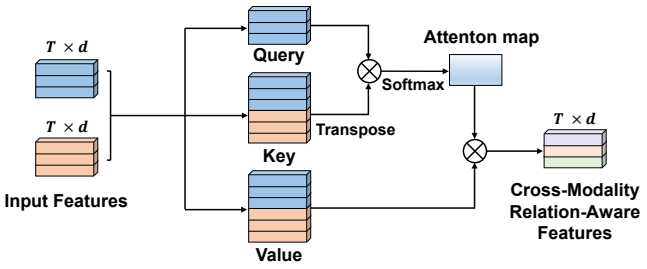

Figure 4: Illustration of the proposed CMRA. The bars in blue and orange represent segment-level features from different modalities. CMRA simultaneously exploits the intra-modality and inter-modality relation information for audio or video segment features.

Without loss of generality, we take the video relation-aware module for illustration. Given visual features $\boldsymbol{v} \in \mathbb{R}^{T \times d_v}$ and audio features $\boldsymbol{a} \in \mathbb{R}^{T \times d_a}$, we first transform them into a common space via linear layers. Here, we denote the transformed visual and audio features as $\boldsymbol{F}_v$ and $\boldsymbol{F}_a$, respectively, with the same dimensions of $T \times d_m$. Then, M_{self} takes as input $\boldsymbol{F}_a$ to explore internal temporal relations in advance, yielding self-attentive audio features denoted as $\boldsymbol{F}_a^s$. Last, M_{cmra} takes as input $\boldsymbol{F}_v$ and $\boldsymbol{F}_a^s$ to explore intra- and inter-modality relations for visual features with the help of CMRA, and yields relation-aware visual features $\boldsymbol{v}_o$ as output. The overall process can be summarized as

$$\begin{aligned} \boldsymbol{v}_o &= M_{cmra}(\boldsymbol{F}_v, \boldsymbol{F}_a^s)\,, \\ \boldsymbol{F}_a^s &= M_{self}(\boldsymbol{F}_a)\,, \\ \boldsymbol{F}_a &= \boldsymbol{a}\boldsymbol{W}_a,\ \boldsymbol{F}_v = \boldsymbol{v}\boldsymbol{W}_v\,, \end{aligned} \tag{4}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{d_a \times d_m}$ and $\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d_m}$ are learnable parameters.

Cross-Modality Relation Attention. Given the audio features and visual features, we aim to exploit cross-modality relations while not neglecting the intra-modality relation information. To this end, we propose a cross-modality relation attention (CMRA) mechanism. We implement CMRA as shown in Figure 4 inspired by the success of self-attention [32].

Specifically, we first project visual features $\boldsymbol{v} \in \mathbb{R}^{T \times d_m}$ into the query features, denoted as $\boldsymbol{q}_1 \in \mathbb{R}^{T \times d_m}$, with a linear transformation. We then temporally concatenate $\boldsymbol{v}$ with $\boldsymbol{a} \in \mathbb{R}^{T \times d_m}$ to obtain a raw memory base $\boldsymbol{m}_{a,v} \in \mathbb{R}^{2*T \times d_m}$. Afterwards, we linearly transform $\boldsymbol{m}_{a,v}$ into key features $\boldsymbol{K}_{1,2} \in \mathbb{R}^{2*T \times d_m}$ and value features $\boldsymbol{V}_{1,2} \in \mathbb{R}^{2*T \times d_m}$. We take the inner-product operation as a pair-wise relation function to measure the intra- and inter-modality relations. A cross-modal attentive output is calculated as

$$\begin{aligned} &\mathbf{CMRA}(\boldsymbol{v}, \boldsymbol{a}) = \mathrm{Softmax}(\frac{\boldsymbol{q}_1(\boldsymbol{K}_{1,2})^T}{\sqrt{d_m}})\boldsymbol{V}_{1,2}\,, \\ &\boldsymbol{q}_1 = \boldsymbol{v}\boldsymbol{W}^Q,\ \boldsymbol{K}_{1,2} = \boldsymbol{m}_{a,v}\boldsymbol{W}^K,\ \boldsymbol{V}_{1,2} = \boldsymbol{m}_{a,v}\boldsymbol{W}^V, \\ &\boldsymbol{m}_{a,v} = \mathrm{Concat}(\boldsymbol{a}, \boldsymbol{v})\,, \end{aligned} \tag{5}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V$ are learnable parameters with dimensions of $d_m \times d_m$ and index 1 or 2 represents different modalities. Note that here we take visual features $\boldsymbol{v}$ as the query for illustration, and we can also take audio features as the query to exploit relations for audio features. In comparison, self-attention can be regarded as a special case of CMRA when the memory contains only the same modality features as the query.

The video/audio relation-aware module in our architecture is a relation-aware module that takes visual/audio features as the query in the CMRA operation.

Cross-Modality Relation Module. Thanks to the CMRA operation, cross-modality relation module M_{cmra} serves to exploit inter-modality relations along with intra-modality relations. Specifically, we first perform CMRA in a multihead setting [32] as

$$\begin{aligned} \boldsymbol{H} &= \mathrm{Concat}(\boldsymbol{h}_1, ..., \boldsymbol{h}_n)\boldsymbol{W}_h\,, \\ \boldsymbol{h}_i &= \mathbf{CMRA}_i(\boldsymbol{F}_v, \boldsymbol{F}_a^s)\,, \end{aligned} \tag{6}$$

where $\boldsymbol{W}_h$ are parameters to be learned, and n denotes the number of parallel CMRA modules. To avoid the transmission loss from CMRA, we add $\boldsymbol{F}_v$ as a residual connection into $\boldsymbol{H}$ along with a layer normalization as

$$\boldsymbol{H}_r = \mathrm{LayerNorm}(\boldsymbol{H} + \boldsymbol{F}_v)\,. \tag{7}$$

To further fuse the information from several parallel CMRA operations, we forward $\boldsymbol{H}_r$ through two linear layers with a ReLU. The detailed calculation of output $\boldsymbol{v}_o$ is given as

$$\begin{aligned} \boldsymbol{v}_o &= \mathrm{LayerNorm}(\boldsymbol{O}_f + \boldsymbol{H}_r)\,, \\ \boldsymbol{O}_f &= \delta(\boldsymbol{H}_r\boldsymbol{W}_3)\boldsymbol{W}_4\,, \end{aligned} \tag{8}$$

where δ denotes the ReLU function, and $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$ are learnable parameters of two linear layers.

Internal Temporal Relation Module. We replace our CMRA with self-attention in M_{cmra} to obtain an internal temporal relation module M_{self}. The module M_{self} concentrates on exploring the internal temporal relation for a portion of memory features in advance to assist in M_{cmra}. We omit the description of M_{self} to avoid repetition.

3.4 Audio-Video Interaction Module

After relation-aware modules, we obtain the cross-modal relation-aware visual and acoustic representations, denoted as $\boldsymbol{v}_o \in \mathbb{R}^{T \times d_m}$ and $\boldsymbol{a}_o \in \mathbb{R}^{T \times d_m}$, respectively. To obtain a comprehensive representation of two modalities for the following classifiers, we propose a simple yet effective audio-video interaction module, which seeks to capture the resonance between visual and acoustic channels by incorporating $\boldsymbol{v}_o$ with $\boldsymbol{a}_o$.

Specifically, we first fuse $\boldsymbol{v}_o$ and $\boldsymbol{a}_o$ with element-wise multiplication to obtain a joint representation of these two modalities, denoted as $\boldsymbol{f}_{av}$. We then leverage $\boldsymbol{f}_{av}$ to attend to the visual representation $\boldsymbol{v}_o$ and acoustic representation $\boldsymbol{a}_o$, where $\boldsymbol{v}_o$ and $\boldsymbol{a}_o$ separately supply visual and acoustic information for better visual understanding and acoustic perception. This operation can be regarded as a variant of our CMRA, where the query is a fusion of the memory features. We denote this operation as $V_{cmra}(\boldsymbol{f}_{av}, (\boldsymbol{a}_o, \boldsymbol{v}_o))$. We then add a residual connection and a layer normalization to the attentive output, similar to the relation-aware module. A comprehensive dual-modality representation $\boldsymbol{O}_{\boldsymbol{av}}$ is calculated as follows:

$$\begin{aligned} \boldsymbol{O}_{av} &= \mathrm{LayerNorm}(\boldsymbol{O} + \boldsymbol{f}_{av})\,, \\ \boldsymbol{O} &= V_{cmra}(\boldsymbol{f}_{av}, (\boldsymbol{a}_o, \boldsymbol{v}_o))\,, \\ \boldsymbol{f}_{av} &= \boldsymbol{a}_o \odot \boldsymbol{v}_o\,, \end{aligned} \tag{9}$$

where $\odot$ denotes element-wise multiplication.

3.5 Supervised and Weakly-Supervised Audio-Visual Event Localization

Supervised Localization. After the audio-video interaction module, we obtain features $\boldsymbol{O}_{av}$ with dimensions of $T \times d_m$. Similar to [36], we decompose the localization into predicting two kinds of scores. One is an event-relevant score $\hat{s}_t$ that determines whether an audio-visual event exists in the t-th video segment. The other one is an event category score $\hat{\boldsymbol{s}}_c \in \mathbb{R}^C$, where C denotes the number of the foreground categories. Event-relevant scores $\hat{\boldsymbol{s}} = [\hat{s}_1, ..., \hat{s}_T] \in \mathbb{R}^T$ are calculated as

$$\hat{\boldsymbol{s}} = \sigma(\boldsymbol{O}_{av}\boldsymbol{W}_s) \,, \tag{10}$$

where $\boldsymbol{W}_s$ are learnable parameters, and σ denotes the sigmoid function. As for the category score $\hat{\boldsymbol{s}}_c$, we conduct max-pooling on the fused features $\boldsymbol{O}_{av}$, yielding a feature vector $\boldsymbol{o}_{av} \in \mathbb{R}^{1\times d_m}$. Afterwards, an event category classifier takes as input $\boldsymbol{o}_{av}$ to predict an event category score $\hat{\boldsymbol{s}}_c$:

$$\hat{\boldsymbol{s}}_c = \text{Softmax}(\boldsymbol{o}_{av}\boldsymbol{W}_c) \,, \tag{11}$$

where $\boldsymbol{W}_c$ is a parameter matrix to be learned.

During the interference phase, the final prediction is determined by $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}_c$. If $\hat{\boldsymbol{s}}_t \geq 0.5$, the t-th segment is predicted to be event-relevant, with an event category according to $\hat{\boldsymbol{s}}_c$. If $\hat{\boldsymbol{s}}_t < 0.5$, the t-th segment is predicted as background.

In the training, we have the segment-level labels, including event-relevant labels and event-category labels. The overall objective function is a summation of a cross-entropy loss for event classification and a binary cross-entropy loss for event-relevant prediction.

Weakly-Supervised Localization. In the weakly-supervised manner, we also predict $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}_c$ as described above. Since we only have access to the video-level labels, we first duplicate $\hat{\boldsymbol{s}}_c$ for T times and $\hat{\boldsymbol{s}}$ for C times, and then fuse them via element-wise multiplication, yielding joint scores $\hat{\boldsymbol{s}}_f \in \mathbb{R}^{T\times C}$. Last, we follow [30] to formulate this problem as an MIL problem [34] and aggregate segment-level predictions $\hat{\boldsymbol{s}}_f$ to obtain a video-level prediction via MIL pooling during training. During inference, the prediction process is the same as that of the supervised task.

4 EXPERIMENT

4.1 Dataset and Evaluation Metric

Dataset. Following previous work [24, 30, 36], we evaluate our method on AVE dataset [30]. It contains 4,143 videos covering a wide scope of domain events (*e.g.*, human activities, animal activities, music performances, and vehicle sounds). The events involve 28 diverse categories (*e.g.*, church bell, baby crying, dog barking, frying food, playing violin, *etc.*). Each video contains one event and is divided into ten one-second segments.

Evaluation Metric. We follow previous work [24, 30, 36] to predict an event category for each segment, and apply overall accuracy as an evaluation metric in the both AVE tasks.

4.2 Implementation Details

Visual Feature Extractor. For fair comparisons, we separately use VGG-19 [29] and ResNet-151 [14] pretrained on ImageNet [22] as visual feature extractors for experiments. Following [30], we uniformly select 16 frames within each segment as input. The output of the *pool*5 layer in VGG-19 with dimensions of $7\times7\times512$ is taken as the visual features. For ResNet-151, we take the output of the *conv*5 layer with dimensions of $7 \times 7 \times 2048$ as visual features. The frame-level features within each segment are temporally averaged as segment-level features.

Audio Feature Extractor. We first transform raw audios into log mel spectrograms and then extract acoustic features with dimensions of 128 for each segment using a VGG-like network [15] pretrained on AudioSet [13].

Training settings. We set the hidden dimension d_m in the relation-aware module as 256. For CMRA and self-attention in relation-aware modules, we set the number of parallel heads as 4. The batch size is 32. We apply Adam [21] as an optimizer. We set the initial learning as 5×10^{-4} and gradually decay it by multiplying by 0.5 at epochs 10, 20 and 30.

Table 1: Comparisons with state-of-the-arts in a supervised manner on AVE dataset

Method	Feature	Accuracy (%)
ED-TCN [23]	VGG-19	46.9
Audio [15]	VGG-like	59.5
Visual [29]	VGG-19	55.3
Audio-Visual [30]	VGG-like, VGG-19	71.4
AVSDN [24]	VGG-like, VGG-19	72.6
Audio-Visual+Att [30]	VGG-like, VGG-19	72.7
DAM [36]	VGG-like, VGG-19	74.5
CMRAN (ours)	VGG-like, VGG-19	**77.4**
Visual [30]	ResNet-151	65.0
Audio-visual [30]	VGG-like, ResNet-151	74.0
Audio-visual+Att [30]	VGG-like, ResNet-151	74.7
AVSDN [24]	VGG-like ResNet-151	75.4
CMRAN (ours)	VGG-like, ResNet-151	**78.3**

4.3 Comparisons with state-of-the-arts

We apply our cross-modal relation-aware networks (CMRAN) in supervised and weakly-supervised AVE localization. For fair comparisons, we compare our method with existing methods using the same features.

Supervised audio-visual event localization. We compare our method with state-of-the-arts using the same visual features as reported in their papers in Table 1. ED-TCN [23] is a state-of-the-art method for temporal action labeling. Our method reaches the highest accuracy using different visual features. Specifically, when using VGG-19 [29], our method achieves 77.44%, surpassing the previous best method by 2.94 % . When using ResNet-151 [14], our method outperforms the previous best method by 2.9 %.

Weakly-supervised audio-visual event localization. We also compare our method with existing weakly-supervised AVE localization methods in Table 2. Our method still achieves the best performance, showing its robustness. Specifically, our methods achieves 72.9% accuracy using VGG-19 visual features, exceeding

Table 2: Comparisons with state-of-the-arts in a weakly-supervised manner on AVE dataset. * indicates the reproduced performance

Method	Feature	Accuracy (%)
AVEL (visual-only) [30]	VGG-19	52.9
AVEL (audio-only)	VGG-like	53.4
AVEL (audio+visual)	VGG-like, VGG-19	63.7
AVEL (audio+visual+ Att)	VGG-like, VGG-19	66.7
AVSDN* [24]	VGG-like, VGG-19	66.8
CMRAN (ours)	VGG-like, VGG-19	**72.9**
AVEL (visual-only)	ResNet-151	63.4
AVEL (audio+visual)	VGG-like, ResNet-151	71.6
AVEL (audio+visual+Att)	VGG-like, ResNet-151	73.3
AVSDN [24]	VGG-like, ResNet-151	74.2
CMRAN (ours)	VGG-like, ResNet-151	**75.3**

the previously best method by 6.1%. When using ResNet-151 visual features, our method achieves the best accuracy of 75.3%.

5 ABLATION STUDIES

In this section, we conduct experiments to verify the effectiveness of each component in the proposed CMRAN. We uniformly use the VGG-19 as a visual feature extractor for experiments.

How does the cross-modality relation attention help? Our CMRA is proposed to exploit useful relation information and further facilitate representation learning. To verify its effectiveness, we implement four variants of CMRAN: "w/o CMRA" , "Self-Att", "Co-Att", and "CMRA-F". These baseline methods are the same as our CMRAN, except that we only replace the CMRA in relation-aware modules with other attention mechanisms or directly remove the CMRA. Specifically, "w/o CMRA" denotes a baseline method where we remove CMRA from relation-aware modules. In the "Self-att" and "Co-att" methods, the CMRA is replaced with self-attention and co-attention respectively. The co-attention here is based on the scaled-dot product attention. The baseline "CMRA-F" equips with a variant of CMRA, where the memory features from two modalities are concatenated along the feature dimension instead of the temporal dimension in CMRA. Furthermore, we try using different linear layers for two modalities of the memory in CMRA, and we did not observe improvements. We argue that this is redundant since we have separately mapped the features of two modalities into a common space at the beginning of relation-aware modules.

As shown in Table 3, the performance of "CMRA-F" significantly declines compared with CMRAN, and this is likely because features from two modalities are entangled together, obstructing a clear relation modeling of two modalities. Critically, the performances of all the variants decrease, verifying the effectiveness of CMRA and justifying the intuition that watching visual contents while hearing the sounds(*i.e.*, capturing intra- and inter-modality relations) is more efficient and effective for localizing an audible and visual event. Besides, both "Self-att" and "Co-att" baselines outperform the "w/o CMRA" baseline, which supports our view that both intra- and inter-modality relations contribute to AVE localization.

Table 3: Ablation study on the effect of CMRA, measured by accuracy(%) on AVE dataset

Method	Supervised	Weakly-supervised
w/o CMRA	76.07	72.03
Self-Att	76.42	72.46
Co-Att	76.64	72.24
CMRA-F	75.62	71.67
with CMRA	**77.44**	**72.94**

Table 4: Ablation study on the effect of AGSCA module and audio-video interaction module, measured by accuracy(%) on AVE dataset

Method	Supervised	Weakly-supervised
w/o AGSCA	76.24	72.16
AGVA [30]	76.95	72.34
AGSCA	**77.44**	**72.94**
w/o AV-Interaction	77.09	72.61
with Self-Att	77.29	72.64
with AV-Interaction	**77.44**	**72.94**

Does the audio-guided spatial-channel attention help? To verify the effectiveness of AGSCA, we implement variants of our CMRAN. Specifically, we remove AGSCA and simply conduct the average pooling spatially and denote this as "w/o AGSCA" method. We also compare our AGSCA with "AGVA" in previous work [30]. As shown in the first part of Table 4, without AGSCA, the performance of our model drops by 1.2% and 0.78% in the supervised and weakly-supervised tasks respectively, which shows that highlighting informative regions enables reduction of background interference for accurate event localization. Comparing our AGSCA with AGVA, the accuracy of AGSCA is higher than that of AGVA, indicating that AGSCA more effectively exploits the audio guidance capability for visual modeling.

Does the audio-video interaction module improve event localization? As shown in the second part of Table 4, we implement two baseline methods, denoted as "w/o AV-Interaction" and "with Self-Att". In the "w/o AV-Interaction" method, we remove the variant of our CMRA from the audio-video interaction module, leaving only an element-wise multiplication. Without the help of the variant of CMRA, the accuracy of event localization decreases absolutely by 0.35% and 0.31% for supervised and weakly-supervised tasks respectively. To further verify the effectiveness, we implement a strong baseline, "with Self-Att", where the variant of CMRA is replaced by self-attention. The performance of this baseline is also lower than that of our method. These experiments imply the good scalability of our CMRA.

Does the internal temporal relation module help? To verify its effectiveness, we first remove the internal temporal relation module (ITR) from relation-aware modules. As shown in Table 5, without ITR, the accuracy drops by 0.71% in the supervised task, indicating that exploiting the internal temporal relation of the other modal representations immediately before performing CMRA enables a performance boost. Besides, we study how the number of

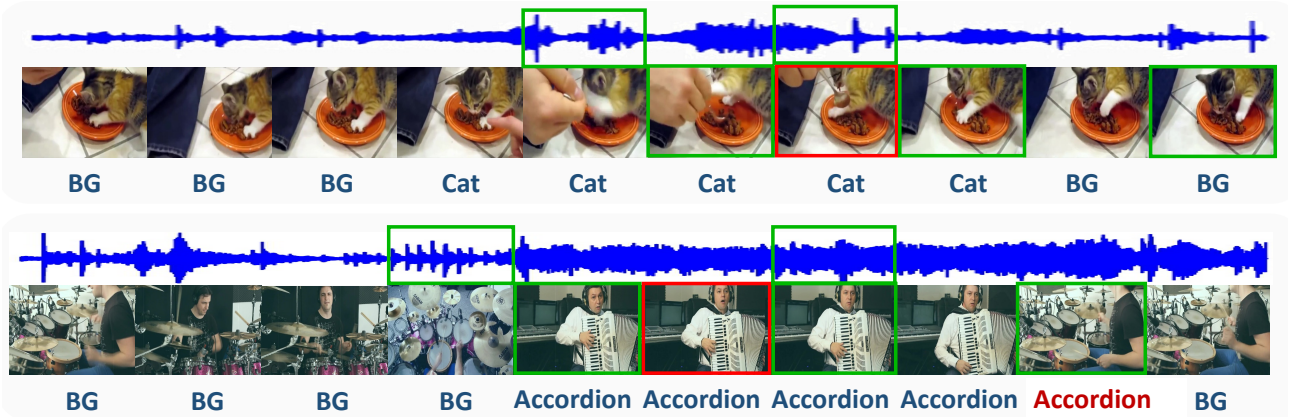

Figure 5: Qualitative results. Our model correctly localizes the cat screaming event in the first example. In the second example, our model fails to correctly predict the category of the ninth segment. The ground truth is "background" while our model predicts it as "Accordion".

Table 5: Ablation study on the effect of internal temporal relation module, measured by accuracy(%) on AVE dataset

Method	Supervised	Weakly-supervised
w/o ITR	76.73	72.56
w/ 1 ITR	77.09	72.91
w/ 2 ITR	**77.44**	**72.94**
w/ 3 ITR	76.86	72.86
w/ 4 ITR	76.67	72.58

ITR modules(L) affects the performance. We found that when $L = 2$, we obtain the best performance in both tasks, and thus we set $L = 2$ in this paper. Furthermore, we also attempted to stack multiple relation-aware modules in CMRAN, but we did not obtain apparent performance improvements. We argue that this may suffer from an overfitting problem.

Figure 6: Qualitative examples of different attention methods. The middle row and bottom row show the visualization examples of AGVA [30] and our method respectively.

5.1 Qualitative Analysis

We show visualization examples of our AGSCA and baseline method AGVA [30] in Figure 6. Owing to the guidance from audio signals for features and spatial attention, our method focuses more on sounding regions (view in color), and covers the sounding regions more accurately and comprehensively (view in the masked area).

We show qualitative examples of our CMRA in Figure 5. The box in red represents a query segment. We mark other segments with top-5 attention weights using green boxes. We found that our CMRA tends to leverage the information of the surrounding segments for event prediction. This is intuitive because the surrounding segments often share similar semantic information that is useful for event prediction, and such information can be regarded as contextual information that can ease the recognition of an AVE. Besides, the segments with high respondence come from two modalities, verifying that our CMRA enables to build connections between two modalities. The second example of Figure 5 contains a failure case. It is challenging since we can hear the sounds of accordion throughout the segment and see an accordion at the very beginning of the segment.

6 CONCLUSION

In this paper, we have devised a relation-aware module that uses the cross-modality relation attention mechanism to capture the useful intra-modality and inter-modality relations for AVE localization. Besides, we have proposed an audio-guided spatial-channel attention module to highlight informative features and spatial regions for reduction of background interference. Built upon these two modules, we present a cross-modal relation-aware network, which significantly outperforms state-of-the-arts in both supervised and weakly-supervised AVE localization tasks on AVE dataset. Empirically, we found some cases where the sounds and visual scenes in a video are not always aligned. It would be interesting to explore our method under this condition, and we leave it for our future work.

ACKNOWLEDGMENTS

This work was partially supported by the Key-Area Research and Development Program of Guangdong Province (2019B010155002, 2018B010108002), National Natural Science Foundation of China (NSFC) (61876208, 61836003 (key project)), Guangdong 2017ZT07X183, Fundamental Research Funds for the Central Universities D2191240, and Pearl River S&T Nova Program of Guangzhou 201806010081.

REFERENCES

[1] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[2] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5659–5667.

[3] Peihao Chen, Chuang Gan, Guangyao Shen, Wenbing Huang, Runhao Zeng, and Mingkui Tan. 2019. Relation Attention for Temporal Action Localization. *IEEE Transactions on Multimedia (TMM)* (2019).

[4] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. 2978–2988.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 4171–4186.

[7] Matthias Fontanellaz, Stergios Christodoulidis, and Stavroula G. Mougiakakou. 2019. Self-Attention and Ingredient-Attention Based Model for Recipe Retrieval from Image Queries. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 25–31.

[8] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. 2020. Foley Music: Learning to Generate Music from Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[9] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. 2020. Music Gesture for Visual Sound Separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*. 2568–2577.

[11] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[12] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. 2020. Listen to Look: Action Recognition by Previewing Audio. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[15] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135.

[16] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3588–3597.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7132–7141.

[18] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware Graph Convolutional Networks for Video Question Answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

[19] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual Attention Networks for Visual Reference Resolution in Visual Dialog. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2024–2033.

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

[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*. 1097–1105.

[23] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 156–165.

[24] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2019. Dual-modality seq2seq network for audio-visual event localization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2002–2006.

[25] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, 689–696.

[26] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7539–7548.

[27] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 801–816.

[28] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh H. McDermott, and Antonio Torralba. 2019. Self-supervised Audio-visual Co-segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2357–2361.

[29] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[30] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[31] Du Tran, Junsong Yuan, and David Forsyth. 2014. Video Event Detection: From Subvolume Localization to Spatiotemporal Path Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36, 2 (2014), 404–416.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*. 5998–6008.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7794–7803.

[34] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3460–3469.

[35] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 2088–2096.

[36] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. 2019. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 6292–6300.

[37] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10502–10511.

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6281–6290.

[39] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. 2019. Breaking Winner-Takes-All: Iterative-Winners-Out Networks for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Image Processing (TIP)* 28, 12 (2019), 5797–5808.

[40] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph Convolutional Networks for Temporal Action Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[41] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Chengwei Zhang, Yunlu Xu, Zhanzhan Cheng, Yi Niu, Shiliang Pu, Fei Wu, and Futai Zou. 2019. Adversarial Seeded Sequence Growing for Weakly-Supervised Temporal Action Localization. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 738–746.

[43] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. 2019. The Sound of Motions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1735–1744.

[44] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh H. McDermott, and Antonio Torralba. 2018. The Sound of Pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 587–604.