Cross-modal Background Suppression for Audio-Visual Event Localization

Yan Xia
Zhejiang University
xiayan.zju@gmail.com

Zhou Zhao*
Zhejiang University
zhaozhou@zju.edu.cn

Abstract

Audiovisual Event (AVE) localization requires the model to jointly localize an event by observing audio and visual information. However, in unconstrained videos, both information types may be inconsistent or suffer from severe background noise. Hence this paper proposes a novel cross-modal background suppression network for AVE task, operating at the time- and event-level, aiming to improve localization performance through suppressing asynchronous audiovisual background frames from the examined events and reducing redundant noise. Specifically, the time-level background suppression scheme forces the audio and visual modality to focus on the related information in the temporal dimension that the opposite modality considers essential, and reduces attention to the segments that the other modal considers as background. The event-level background suppression scheme uses the class activation sequences predicted by audio and visual modalities to control the final event category prediction, which can effectively suppress noise events occurring accidentally in a single modality. Furthermore, we introduce a cross-modal gated attention scheme to extract relevant visual regions from complex scenes exploiting both global visual and audio signals. Extensive experiments show our method outperforms the state-of-the-art methods by a large margin in both supervised and weakly supervised AVE settings. [1]

1. Introduction

Event location and action recognition [3, 10, 30] have become increasingly important in understanding and analyzing video content, with most methods relying on optical flow and RGB features processing. However, audio can also provide valuable clues for understanding holistic video content [11, 22, 37]. To comprehensively realize how to combine audio and visual modalities and understand the video contents, Tian et al. [32] introduce the audio-visual event (AVE) localization task, where the model determines the presence of an event and localizes its boundary in the temporal dimension when the event is both audible and visible at the same time. Figure 1 demonstrates that this way of judging an audio-visual event is necessary, as only simultaneously observing the car and hearing the engine sound indicates that the car is running. Compared with the traditional video event localization, the challenges of the AVE task mainly exist in the following aspects [1, 21, 38]: i). Merging the complementary audio and visual features while preserving the simultaneously modal-specific information is not trivial. ii). Sudden noise and complex background existing in the unconstrained videos will hinder the predictions of the event categories. iii). AVE requires the event to be audible and visible, while unsynchronized audio and visual information mislead the event boundary prediction.

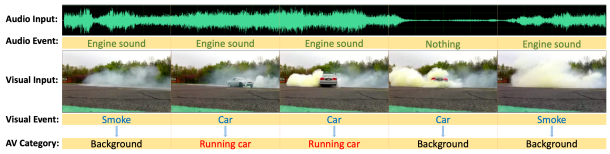

Figure 1. An illustration example of AVE task. An audio-visual event will be identified when it is both audible and visible. In this example, only when we see the car and hear the engine sound at the same time (the 2nd and 3rd segments) can we determine an audio-visual event *"running car"* happened.

Early models mainly focused on solving the first challenge by simply fusing the information of the two modalities after processing each modality independently [14,32] or aligning audio and visual information and then fusing them by cross attention [4, 34, 36]. However, problems such as event category detection errors caused by the sudden noise existing in single mode or inaccurate temporal event localization caused by unsynchronized audio and visual information are still an open research case.

Unlike previous methods, we consider the problem of audio-visual event localization from the viewpoint of cross-modal background suppression, which can effectively alleviate challenges (ii) and (iii). Background suppression methods have been successfully used in the previous weakly

*Corresponding author.
[1]The source code and pre-trained models are publicly available at: https://github.com/marmot-xy/CMBS

supervised temporal action localization [13, 23, 25], by designing a background-aware attention module to distinguish the foreground from background in the visual modality. However, the above methods can only determine whether a video segment belongs to background or not in single modality (i.e, audio or vision). While in AVE task, a video segment may be regarded as foreground by these methods in one mode, but in fact it may be background because relevant information is missing in the other modality [14, 34]. Additionally, several ambiguous and sustained audio or visual noises exist in unconstrained videos [35, 36], which can not be distinguished well by the traditional background suppression methods employing a single modality [13, 20]. Thus, it is necessary to develop a new cross-modal background suppression model to alleviate the existing problems in multi-modal tasks.

In this paper, we first define the "background" category from two aspects: 1) If the audio and visual information in the small video segment do not represent the same event, then the video segment will be labeled as background. 2) If an event only occurs in one modality but has a low probability in another, then this event category will be labeled as background in this video, i.e., offscreen voice. Hence, this paper proposes a novel cross-modal background suppression method considering two aspects: time-level and event-level, which allow the audio and visual modalities to serve as the supervisory signals complementing each other to solve the aforementioned AVE task problems. The time-level background suppression can convert the audio and visual features to temporal dimension gates and make the model reduce its attention to the ambiguous video segments that only one modal information is meaningful. As for challenge ii), we design an event-level background suppression method, which considers that a low probability event in just one modality should not appear in the another modality. In this way, even if noise appears in the audio or vision but is absent in another mode, the final recognition result can effectively suppress these noises. Furthermore, to avoid the interference of some unimportant background objects in the video, we propose a novel cross-modal gated attention (CMGA) module, which can use audio and visual global information to do cross-modal gating to jointly control the feature extraction of useful visual regions. We integrate these modules and propose a Cross-Modal Background Suppression model, which outperforms the current state-of-the-arts methods by a large margin in both supervised and weakly supervised AVE settings.

2. Related Work

2.1. Audio-Visual Representation Learning

In recent years, many works have been explored to learn audio-visual representation learning, which can be divided into two main classes according to the supervision method. Some methods focused on learning the fusion of audio and visual information with the supervised signals, like [7, 11, 16, 18, 19, 24]. Kazakos et al. [11] proved that using a late-fusion of audio and visual modalities will achieve better performance than fusion before temporal aggregation. Nawaz et al. [24] used a single stream network to jointly embedding the audio and visual features to a shared latent space without pairwise or triplet information. Long et al. [16] designed four different multi-modal fusion methods to find which is better for discerning interactions between modalities, such as future fusion, LSTM fusion, attention fusion and probability fusion.

Many other works focused on how to investigate the cross modal representation with unsupervised or contrastive learning methods [17, 21, 21]. Early works like [2] and [26] learned such a representation used a neural network to predict whether video frames and audio are temporally aligned. Hu et al. [8] proposed a Deep Multi-modal Clustering (DMC) network to perform elaborate correspondence learning among audio and visual components. Alwassel et al. [1] proposed a Cross-Modal Deep Clustering (XDC) to leverage unsupervised clustering in one modality (e.g., audio) as a supervisory signal for the other modality (e.g., video), which can utilize the semantic correlation and the differences between the two modalities. The same idea like Zhang et al. [37] used the knowledge shared between audio and visual modality serves as a supervisory signal.

2.2. Audio-Visual Event Localization

Tian et al. [32] first brought up an audio-visual event localization dataset and treated the AVE problem as a sequence labeling problem. They proposed a dual multimodel residual network to fuse information over the two modalities. AVSDN [14] utilized both audio and visual data at each time segment as inputs and exploited global and local event information in a seq2seq manner. These methods directly contacted the two features at the segment level, which can cause the content of the two modalities to be misaligned temporally, thus Wu et al. [34] proposed a dual attention matching module which first captured the global event-level information for each modality and then checked segment-level local temporal information by a global cross-check mechanism. Duan et al. [4] first applied the co-attention mechanism in AVE task. Xuan et al [36] designed a cross model attention network to extract "where", "when" and in "which" sensor the most related event information between audio and video. Lin et al. [15] proposed an audiovisual transformer network which can jointly encode intra-frame and inter-frame visual features by observing audio features. Xu et al. [35] proposed an audio-guided spatial-channel attention which can guide the model to focus on event-relevant visual regions. To ignore the interfer-

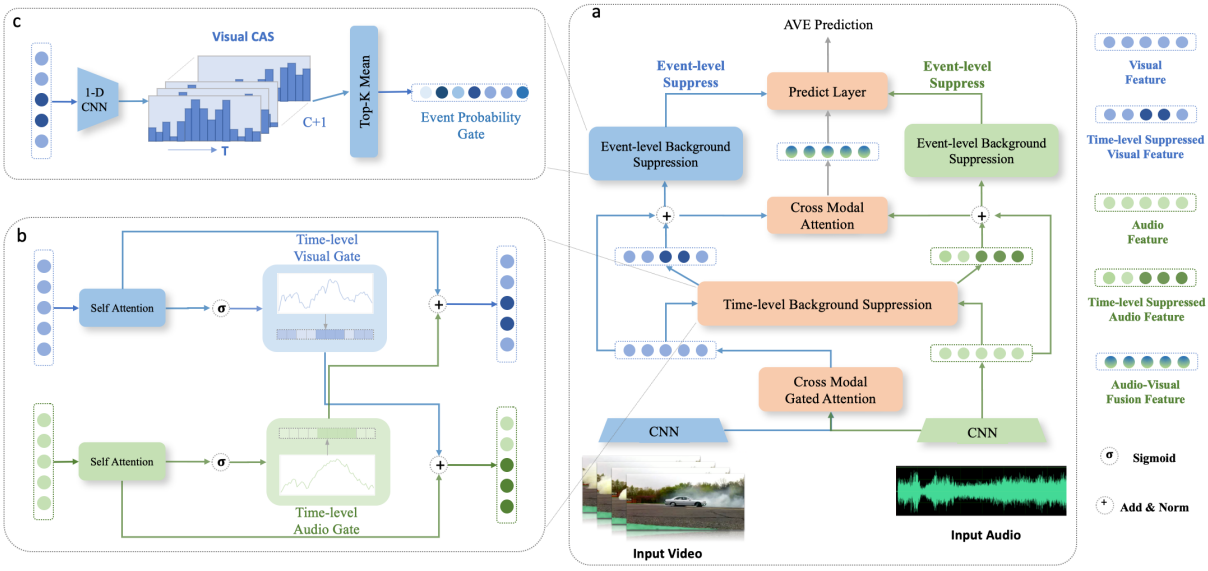

Figure 2. The proposed cross-modal background suppression network. (a) The main pipeline of our model. (b) We utilize a time-level background suppression to suppress the unimportant information in the other modality according to the importance of each modality content. (d) We leverage audio and visual information and apply an event-level background suppression scheme to suppress the events occurring with low probability in a single modality.

ence caused by irrelevant audio-visual segment pairs, Zhou et al. [38] proposed a positive sample propagation (PSP) module which is able to select positive connections of two modalities and ignore the negative connections. Tian et al. [31] developed a hybrid attention network to better extract temporal contexts simultaneously from unimodal and cross-modal, then use an attentive MMIL pooling method to adaptively explore useful audio and visual content from two modalities.

The above methods focused on how to better integrate the audio and visual information together, while ignoring the potential supervisory information of the audio and visual modality for each other. Different from these models, in this paper, we propose cross-modal time-level background suppression to force audio and visual modalities to pay more attention to the segments that both modalities consider important. Also we bring up a cross-modal event-level background suppression which can force the model to ignore the occurrence of noise in single modal and predict event categories more accurately.

3. Methods

3.1. Problem Setting

First we present some notations used in the subsequent illustrations and then introduce the problem of audio-visual event localization from the viewpoint of cross-modal background suppression. In general, the AVE task is defined as follows: given a video sequence with T non-overlapping segments $S = (A_t, V_t)_{t=1}^{T}$, where each segment is one second, the model is required to predict the event label (including background) of each segment as $y_t = \{y_t^k | y_t^k \in \{0,1\}, k = 1, ..., C, \sum_{k=1}^{C} y_t^k = 1\}$ during the inference phase; where $A_t \in R^{d_a}$ and $V_t \in R^{H \times W \times d_v}$ denote the corresponding audio and visual features extracted by the pre-trained models of the t-th segment, d_a and d_v are the audio and visual dimensions, H and W are the height and width of the visual feature map, and C is the number of the categories (including the background). $y_t^k = 1$ refers to t-th video segment, where the k-th event is both audible and visible. Otherwise, it will be zero. This paper mainly studies the AVE task in two different settings, which are divergent for given ground truth labels. In **fully-supervised Setting**, the segment-labels $Y^f \in R^{T \times C}$ are available during the training phase, indicating whether the video segment contains an event and its category type. While in **weakly-supervised Setting**, only the video-labels $Y^w \in R^{1 \times C}$ are available during the training phase, which are the average pooling of the full ground truth labels.

3.2. Overall Pipeline

Figure 2 illustrates the architecture of the proposed method. First, in unconstrained videos, many complex and redundant visual backgrounds may occur that are irrelevant to the task, i.e., the smoke and trees as shown in Figure 1. To bridge this gap, we propose a cross-modal gated attention module which guides the audio and global visual signals to select the informative visual regions and reduce the irrelevant background interference from both channels.

Next, we devise a cross-modal time-level background

suppression module to determine the video segments on which audio and visual modalities should focus. Subsequently, we use cross-modal attention and a residual layer to integrate relevant information in the audio and video modalities. Finally, we utilize the audio-video interaction module proposed in [35] and a classification layer to predict the final event labels. At the same time, we use event-level background suppression as two gates to control the event predictions. These modules will be introduced in detail in the following sections.

3.3. Cross-Modal Gated Attention

Previous works like AGVA [32] and AGSCA [35] utilized audio signals to guide the visual features extraction, which are effective for most situations. However, harsh noise embedded in the audio signal greatly hinders the extraction of essential vision information. Inspired by the great success of non-local block [33] and SENet [9] for visual tasks, this paper utilizes both global visual signals and audio signals to develop a cross-modal gated attention module that robustly extracts the event-relevant visual regions.

We first perform channel-level attention on visual features with both audio signals a_t and global visual signals v_t^g as described in [9]. After obtaining the channel-level attention features v_{gt}^c and $v_{at}^c \in R^{(HW)\times d_v}$, we then perform spatial-level attention to extract the important visual spatial information from v_{gt}^c and v_{at}^c. The spatial level attention scores for v_t^g and a_t are represented as A_{gt}^s and A_{at}^s. For example, the details for calculating A_{at}^s can be formulated as follows:

$$\begin{aligned} u_t^a = \delta(W_a a_t), u_{gt}^c = \delta(W_v v_{gt}^c), \\ A_{gt}^s = softmax(\delta(W_1(u_t^a \odot u_{gt}^c))), \end{aligned} \tag{1}$$

where W_a, W_v and W_1 are three learning parameters, $\odot$ is the Hadamard product and δ is the Relu activation function. Then, we multiply A_{gt}^s and A_{at}^s with v_{gt}^c and v_{at}^c, respectively, to obtain the spatial-level attention visual features as v_{gt}^s and $v_{at}^s \in R^{d_v}$. Finally, we design a cross-gated mechanism to select the important visual regions of t-th segment with two residual gates σ_a and σ_g produced from A_{gt}^s and A_{at}^s:

$$V_t = \frac{1}{2}(v_{gt}^s + \beta_1 * \sigma_a v_{gt}^c) + \frac{1}{2}(v_{at}^s + \beta_2 * \sigma_g v_{at}^c), \tag{2}$$

where β_1 and β_2 are hyperparameters. σ_a and σ_g can select the event-relevant regions considered by audio signals and global visual signals to supplement spatial-level attention visual features. The final visual feature vector $V \in R^{T\times d_v}$ obtained contains the channel-spatial attentive information from both the audio and global visual memory features, which can reduce the irrelevant background features and improve the quality of the visual representations.

3.4. Time-level Background Suppression

This section describes the details on the cross-modal time-level background suppression. Previous works [13, 23, 25] proved that background suppression is helpful in only one modality, while in this paper, after obtaining the features of audio and visual modalities, we leverage the information from both modalities as supervised signals to distinguish and suppress the ambiguous backgrounds for each other. To be specific, we first apply a self-attention mechanism on these two modalities in the time dimension and obtain $V_s, A_s \in R^{T\times d_h}$, where d_h is the hidden dimension. Each segment of V_s and A_s contains audio or visual related information of the entire video, helping to understand the overall content in the time level. To determine the important segments for each modality, we add two gates for V_s and A_s, respectively: $g_v^{tbs} = \sigma(W_g^v \cdot V_s)$ and $g_a^{tbs} = \sigma(W_g^a \cdot A_s)$, where $W_g^v, W_g^a \in R^{d_h\times 1}$. The values in both gates represent how important each segment is considered by the audio and visual modality. For example, if the visual modality considers a certain segment containing the main event, the value of the corresponding position of g_v^{tbs} will be larger. Then we multiply the visual gate g_v^{tbs} with the audio features A_s, and the audio gate g_a^{tbs} with the visual features V_s to select the fragments that these two gates consider important from the opposite modalities. Finally, we exploit the residual connection idea that multiplies two modality features by a coefficient and obtain the time-level suppressed visual and audio features V_s^{tbs} and A_s^{tbs}, respectively. The latter features are formulated as follows:

$$V_s^{tbs} = (1-\alpha)V_s + \alpha * (g_a^{tbs} \cdot V_s), \tag{3}$$

$$A_s^{tbs} = (1-\alpha)A_s + \alpha * (g_v^{tbs} \cdot A_s), \tag{4}$$

where the α is a hyperparameter. By enhancing the information of the important parts of the video fragment, the information of the background parts will be indirectly suppressed. Here we only weaken the attention to the background information, but we do not completely erase these information because it also plays a great role in understanding the content of the entire video. In the following ablation studies, we also prove through experiments that appropriately weakening these background information can improve the model's ability to locate the event, while excessive suppression is harmful.

3.5. Event-level Background Suppression

After suppressing background at the temporal level, we obtain the time-level suppressed audio and visual features. Nevertheless, some mixed sounds in audio or complex scenes in vision can not be distinguished well when employing a single modality, impacting the model's prediction for event categories. To alleviate this problem, we propose a cross-modal event-level background suppression scheme

that exploits both audio and visual information and suppresses noise events with low probability for each other. Specifically, we first load the time-level suppressed visual and audio features V_s^{tbs} and A_s^{tbs} into the temporal 1D convolutional layers and predict the segment-level classification scores S_V and S_A using Class Activation Sequences (CAS).

For fully-supervised and weakly-supervised tasks, we have different dimensions for S_V and S_A: for fully-supervised settings, $S_V, S_A \in R^{T\times C}$, while for weakly-supervised settings, $S_V, S_A \in R^{T\times(C+1)}$, and C is the event category number. The reason for the different dimensionality will be described in detail later. Next, we aggregate the segment-level event scores using class-wise top-K mean technique [27] to get $\hat{S_V}$ and $\hat{S_A}$:

$$\hat{S_m} = \frac{1}{K} \underset{s_{n;c}\in S_m[c,:]}{max} \sum_{\forall s\in s_{n;c}} s, \tag{5}$$

where $m \in \{A, V\}$, $s_{n;c}$ is a subset containing the top-K event scores for the c-th class and K is a hyperparameter controlling the number of the selected segments in a video, commonly set to 4. Finally we apply the sigmoid activation function on $\hat{S_V}$ and $\hat{S_A}$ to obtain the visual and audio event-level gates: $g_v^{ebs} = \sigma(\hat{S_V})$ and $g_a^{ebs} = \sigma(\hat{S_A})$. The g_v^{ebs} and g_a^{ebs} contain the probability of each event occurring in the audio and visual modalities and will be multiplied with the final prediction result in the subsequent process. The primary function of these two gates is to suppress events that occur in only one modality but have a low probability in another modality, which can decouple the mixed noise that cannot be distinguished in a single mode and improve the robustness of the final predictions.

3.6. Classification and Objective Function

Before fusing audio and vision information, we employ cross-modal attention to afford each modality exploiting relevant information from the other modality. Previous works [35] have proved that connecting the audio and visual features as key and value matrices is useful. Adopting this approach, we use the multi-head attention, and residual connection mechanism to fuse the valuable information for time-level suppressed audio and visual features. In this scheme, the query features $Q \in R^{T\times d_m}$ are audio or visual matrices, while the key and value feature $K, V \in R^{2\times T\times d_m}$ are the concatenation of audio and visual matrices in the time dimension. This process provides the visual and audio modality matrices denoted as V_o and A_o, respectively. After that, we use the audio-visual interaction module proposed in [35] to obtain a comprehensive fusion of audio and visual information as F_o for the following classifier.

Fully-supervised AVE task: Following [34], we decouple the supervised audio-visual event localization task into two subtasks: one involves predicting an event category label as $\hat{y}_c \in R^C$ and the other to predict an event-relevant score $\hat{y}_t \in R^T$ that judges whether the audio and visual events in t-th video segment are consistent. We also apply the audio and visual event-level background gates g_v^{ebs} and g_a^{ebs} to suppress the noisy events existing in $\hat{y}_c$. Specifically, the $\hat{y}_c$ and $\hat{y}_t$ can be calculated as:

$$\hat{y_t} = softmax(W_{ot}F_o), \tag{6}$$

$$\bar{y_c} = W_{oc}\bar{F}_o, \tag{7}$$

$$\hat{y_c} = (1-\gamma)\bar{y_c} + \gamma * (g_v^{ebs} \odot g_a^{ebs}) * \bar{y_c}, \tag{8}$$

where $W_{ot} \in R^{d_h\times 1}$, $W_{oc} \in R^{d_h\times C}$, $\bar{F}_o$ is the max-pooling result from F_o, and γ is a hyperparameter. It should be noted that we determine whether t-th video segment belongs to event or background, according to the value of $\hat{y}_t$. Thus we only need to predict the event category number as C and not C+1. This explains why we predict the dimension of CAS scores S_v and S_a for fully-supervised settings as $R^{T\times C}$. Additionally, to optimize the time-level background gates accordingly, we multiply the audio gate g_a^{tbs} with the visual gate g_v^{tbs} to obtain the audio-visual gate g_{av}^{tbs}, which is optimized by calculating loss based on the ground truth (GT) event relevance label. During training, we obtain both the corresponding GT event category and the relevance label in fully-supervised settings. Thus the overall objective function is:

$$L_{fully} = L^c + \frac{1}{N}\sum_{t=1}^{N}(L_t^r + L_t^g), \tag{9}$$

where L^c is the cross-entropy loss of the event category between prediction $\hat{y}_c$ and GT label, L_t^r refers to the binary cross entropy loss of the event relevance between the prediction $\hat{y}_t$ and t-th segment GT label, and L_t^g denotes the binary cross entropy loss between gate g_{av}^{tbs} and the t-th segment GT event relevance label. More detailed discussion on the L_t^g function is provided in the ablation study section. During the inference phase, if $\hat{y}_t > 0.5$ then the t-th video segment is predicted as $\hat{y}_c$ class. Otherwise the t-th video segment is classified as background.

Weakly-supervised AVE task: For an AVE task, we adopt [32] and formulate the weakly supervised problem as multiple-instance learning (MIL) problem. Since only the event category label is available during training, we do not predict the event-relevant score $\hat{y}_t$. Furthermore, since we cannot utilize the GT time-level event relevance label to optimize our predicted time-level audio-visual gate, we duplicate the AV gate g_{av}^{tbs} for C+1 times and $\bar{y_c}$ for T times and use element-wise multiplication to fuse them. Then we exploit [32] and utilize MIL pooling to aggregate the results and obtain the video-level event predictions. This process is

formulated as follows:

$$\bar{y_c} = W_{oc} F_o, \tag{10}$$

$$\bar{y_c} = (1-\gamma)\bar{y_c} + \gamma * (g_v^{ebs} \odot g_a^{ebs}) * \bar{y_c}, \tag{11}$$

$$\hat{y_c} = \text{max-pooling}(g_{av}^{tbs} \odot \bar{y_c}), \tag{12}$$

where $W_{oc} \in R^{d_h \times (C+1)}$. In weakly supervised settings, we need to predict the event category label and the background simultaneously. Therefore, the event category number is C+1. The objective function is the a softmax function and multi-class cross-entropy loss, while the loss function indirectly optimizes the time-level audio-visual gate.

4. Experiments

This section presents the experimental setup details and then challenges our methods against state-of-the-art methods on the AVE dataset under two settings. We also discuss the model's performance and specify the effectiveness of each sub-module in our model through ablation studies and qualitative results.

4.1. Experiment Setup

AVE Dataset: Tian et al. [32] created the AVE dataset originating from the AudioSet [5], which contains 4,143 videos covering 28 event categories. The AVE dataset involves a large variety of videos, such as church bell, race car, women or men speaking, and dog barking. Each video lasts 10 seconds and contains an event. The audio-visual event categories are labeled for each video on a segment level.

Evaluation Metrics: For the AVE task, the event category label of each video segment is required to be predicted in both supervised and weakly supervised settings. We adopt [32, 36] and exploit the overall accuracy of the predicted event category as the evaluation metric.

Implementation Details: Regarding the AVE dataset, we exploit the pre-trained on ImageNet [12] VGG-19 and ResNet-151 models to extract visual features of size $7 \times 7 \times 512$ and $7 \times 7 \times 2048$, respectively, per video segment. We utilize a VGG-like network [6] pre-trained on AudioSet to extract 128-dimensional audio features per audio segment. We use a batch size of 64, and the optimizer is Adam. The learning rate is 7×10^{-4} and gradually decays to 0.5 at epochs 10, 20, 30 [35].

4.2. Comparisons with State-of-the art Methods

We challenge our method against current fully-supervised and weakly-supervised methods on AVE tasks. For a fair comparison, we choose the same audio and visual features as the current methods. Specifically, since current techniques utilize two different visual features of VGG-19 and Res-151 in the AVE task, we also apply these two features to our model and compare them with the previous methods.

Table 1. Comparisons with state-of-the-art methods in a supervised manner on the AVE dataset

Models	Feature	Accuracy(%)
Audio	VGG-like	59.5
Visual	VGG-19	66.1
AV [32]	VGG-like, VGG-19	71.4
AVSDN [14]	VGG-like, VGG-19	72.6
CMAN [36]	VGG-like, VGG-19	73.3
DAM [34]	VGG-like, VGG-19	74.5
AVRB [29]	VGG-like, VGG-19	74.8
AVIN [28]	VGG-like, VGG-19	75.2
AVT [15]	VGG-like, VGG-19	76.8
CMRAN [35]	VGG-like, VGG-19	77.4
PSP [38]	VGG-like, VGG-19	77.8
Ours	VGG-like, VGG-19	**79.3**
Visual [32]	Res-151	65.0
AV [32]	VGG-like, Res-151	74.0
AVSDN [14]	VGG-like, Res-151	75.4
CMRAN [35]	VGG-like, Res-151	78.3
Ours	VGG-like, Res-151	**79.7**

Table 2. Comparisons with state-of-the-art methods in a weakly supervised manner on the AVE dataset

Models	Feature	Accuracy(%)
AVEL [32]	VGG-like, VGG-19	66.7
AVSDN [14]	VGG-like, VGG-19	67.3
CMAN [36]	VGG-like, VGG-19	70.4
AVRB [29]	VGG-like, VGG-19	68.9
AVIN [28]	VGG-like, VGG-19	69.4
AVT [15]	VGG-like, VGG-19	70.2
CMRA [35]	VGG-like, VGG-19	72.9
PSP [38]	VGG-like, VGG-19	73.5
Ours	VGG-like, VGG-19	**74.2**
AVEL [32]	VGG-like, Res-151	71.6
AVSDN [14]	VGG-like, Res-151	74.2
CMRAN [35]	VGG-like, Res-151	75.3
Ours	VGG-like, Res-151	**76.0**

Supervised localization for AVE. Table 1 demonstrates our method's performance against current state-of-the-art methods on supervised AVE tasks. Most mainstream models adopt VGG-like audio features and VGG-19 visual features as their input. Compared with them, our method achieves a new state-of-the-art (SOTA) performance. Notably, our method significantly outperforms CMRAN [35] and PSP [38] by a large margin (1.9% and 1.5%, respectively), demonstrating the effectiveness of our proposed method. Even when we use VGG-19 visual features while the competitor methods adopt Res-151, our model still outperforms them. Furthermore, when we utilize Res-151, our technique surpasses the previous SOTA CMRAN method

Table 3. Ablation studies of different modules on AVE dataset.

Models	Supervised	Weakly-supervised
full model	79.30	74.22
w/o CMGA	78.07	71.91
w/o Self-Att	77.62	71.29
w/o TBS	77.93	73.45
w/o Cross-Att	78.10	72.03
w/o EBS	78.30	73.57

[35] by 1.4%.

Weakly-supervised localization for AVE. We compare our method with existing weakly-supervised AVE localization state-of-the-arts. As presented in Table 2, our method still achieves the best performance, demonstrating our model's effectiveness in both tasks. Specifically, our model outperforms CMRAN [35] and PSP [38] by 1.3% and 0.7%, respectively, when using the VGG-19 visual feature, and outperforms CMRAN [35] by 0.7% when using the Res-151 visual feature. Moreover, with a much lower supervision level, our model outperforms some early fully-supervised methods.

4.3. Ablation Studies

The effectiveness of the different modules in our model. To verify the effectiveness of the proposed modules, we remove them from the primary model and re-evaluate the new model on both two tasks. Table 3 shows that after removing a single module, the overall model's performance decreases, and different modules have different performance effects. Self-attention and cross-attention integrate the audio and visual inter and intra-modality information. Thus the accuracy scores will severely drop when one of these modules is removed. Our proposed CMGA, time-level background suppression (TBS), and event-level background suppression (EBS) also play an essential role in the event localization accuracy.

Influence of the time-level background suppression hyper-parameter α. To further explore the effectiveness of our time-level background suppression (TBS), we assign different values to α to observe the model's performance on the supervised scheme, with the corresponding results presented in Table 4. Specifically, $\alpha = 0$ refers to only adding an audio-visual gate loss L_t^g to the final objective function for model training, without suppressing the time-level audio or visual features. Without TBS means that we also need to remove the L_t^g. Compared to the accuracy of neglecting the TBS module, the L_t^g can improve our model's performance. When α is small, the model's performance gradually improves as α becomes larger. However, when α exceeds 0.2, the model's performance gradually declines, inferring that excessive suppression of the audio and video features in specific segments reduces the model's performance due to information discontinuity. Hence, proper suppression is mandatory to achieve better performance.

Table 4. Ablation study of the time-level background suppression (TBS) in supervised manner on AVE dataset.

Models	Accuracy(%)
with TBS $\alpha = 0$	78.37
with TBS $\alpha = 0.1$	78.79
with TBS $\alpha = 0.2$	**79.30**
with TBS $\alpha = 0.3$	78.45
with TBS $\alpha = 0.4$	78.58
w/o TBS	77.93

Table 5. Ablation studies on the effect of event-level background suppression on AVE dataset.

Models	Supervised	Weakly-supervised
full model	79.30	74.22
w/o audio EBS	79.05	73.83
w/o visual EBS	78.84	73.40
w/o EBS	78.30	73.57

Impact of event-level background suppression. To verify the effectiveness of the event-level background suppression (EBS), we test the performance relying on a single EBS modality and compare the results against the entire model. As illustrated in Table 5, both modalities effectively suppress ambiguous events in the opposing modality, and removing the visual EBS imposes a minor performance drop compared with removing the audio EBS. Hence, this illustrates the importance of using visual information to reduce the noise existing in the audio modality.

4.4. Qualitative analysis

Figure 3 and 4 present qualitative examples of the effectiveness of our cross-modal time-level background suppression. These two examples are very representative. Fig 3 refers to visual information being visible most of the time. However, the relevant audio event only appears in a part of the period. Fig 4 refers to audio information being audible during the entire process while the relevant visual event is only available in part of the segments. By observing the audio and visual gate values, we can find that the time-level background suppression scheme suppresses well the attention to the segments when only single modal information is available. This allows the model to better judge the event boundary and confirms that audio can distinguish some ambiguous visual actions mentioned earlier. For example, in Fig 3, the dog's posture hardly changes, except that there is an inconspicuous mouth opening action in the second and third segments. It is challenging to perform recognition solely relying on visual information. However, our model can quickly know that the second and third segment infor-

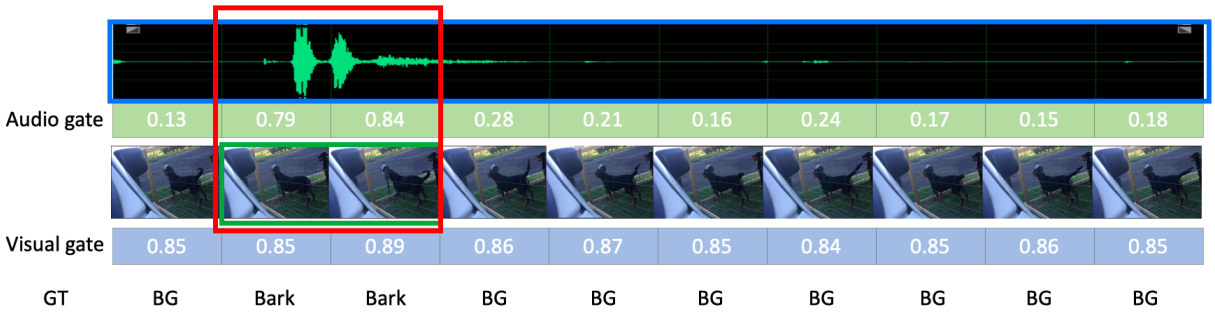

Figure 3. Qualitative results of our model on *dog bark* event. The red regions stand for the answer we predict.

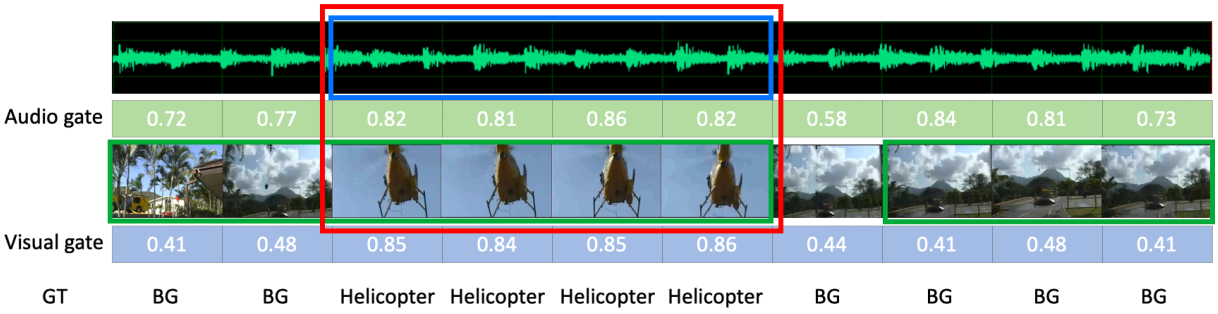

Figure 4. Qualitative results of our model on *helicopter* event. The red regions stand for the answer we predict.

Figure 5. The qualitative results of our proposed cross-modal gated attention. The top row shows the event category, while the middle and bottom rows show the original video frames and visualization of the attention maps, respectively.

mation is of greater interest in the visual modality when audio is added.

We show the visualization results of our cross-modal gated attention module. From Figure 5 we observe that with the guidance of the audio signals and global visual information, the module can accurately focus on the critical visual regions in those video frames for different events. For example, in the example of the banjo event, 4-5 banjos exist in the frame, and our model exactly localizes the two sounding banjos while ignoring the other instruments. Global visual information also plays a vital role in the extraction of critical visual features. In the sixth example of the bark event, the dog is barking very weakly. Thus, only with the guidance of audio the model incorrectly extract the key visual features. However, with the help of global visual information that assists in understanding the holistic content of the video frames, the model can still focus on the suitable regions.

5. Conclusion

This paper proposes a cross-modal time-level and event-level background suppression to better solve the problem of inconsistent audio and visual information within an AVE task. Our purpose is to exploit the audio and visual information as a supervisory signal for the opposite modality, which can help the model focus on the video segments when the events are audible and visible. Also, the event-level background suppression can utilize the CAS scores predicted by audio and vision to suppress the events with low probability in one modality. The experiments demonstrate that EBS effectively suppresses the noise events that suddenly occur in a single-mode, improving the model's robustness. Besides, we also devise a cross-modal gated attention module to better extract the key visual region features from the complicated video frames by exploiting audio and global visual information guidance. Extensive experiments demonstrate that our proposed CMBS network outperforms current state-of-the-art methods in both fully-supervised and weakly-supervised AVE tasks. Also, the ablation studies on these datasets verify the effectiveness of our proposed background suppression methods and the CMGA module.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grant No.62072397, Zhejiang Natural Science Foundation under Grant LR19F020006.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS 2020,*, 2020. 1, 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV 2017, Venice,*, pages 609–617. IEEE Computer Society, 2017. 2

[3] Chenglizhao Chen, Guotao Wang, Chong Peng, Yuming Fang, Dingwen Zhang, and Hong Qin. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. *IEEE Transactions on Image Processing*, 30:3995–4007, 2021. 1

[4] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV 2021*, pages 4012–4021. IEEE, 2021. 1, 2

[5] Jort F. Gemmeke, Daniel P. W. 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,*, pages 776–780. IEEE, 2017. 6

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 ICASSP*, pages 131–135. IEEE, 2017. 6

[7] Chiori Hori, Takaaki Hori, Gordon Wichern, Jue Wang, Teng-Yok Lee, Anoop Cherian, and Tim K. Marks. Multimodal attention for fusion of audio and spatiotemporal features for video description. In *CVPR Workshops 2018*, pages 2528–2531. IEEE Computer Society. 2

[8] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR 2019*, pages 9248–9257. Computer Vision Foundation / IEEE, 2019. 2

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR 2018*, pages 7132–7141. IEEE Computer Society, 2018. 4

[10] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. Video moment localization via deep cross-modal hashing. *IEEE Transactions on Image Processing*, 30:4667–4677, 2021. 1

[11] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV ,2019*, pages 5491–5500. IEEE, 2019. 1, 2

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. 6

[13] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI 2020,*. 2, 4

[14] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019*. IEEE, 2019. 1, 2, 6

[15] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Computer Vision - ACCV 2020*, volume 12627 of *Lecture Notes in Computer Science*, pages 274–290. Springer, 2020. 2, 6

[16] Xiang Long, Chuang Gan, Gerard de Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *(AAAI-18)*, pages 7202–7209. AAAI Press, 2018. 2

[17] Shuang Ma, Zhaoyang Zeng, Daniel J. McDuff, and Yale Song. Learning audio-visual representations with active contrastive coding. *CoRR*, abs/2009.09805, 2020. 2

[18] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylène C. Q. Farias, and Alan Conrad Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29:6054–6068, 2020. 2

[19] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020. 2

[20] Md. Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C. Leu. Action completeness modeling with background aware networks for weakly-supervised temporal action localization. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2166–2174. ACM, 2020. 2

[21] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. *CoRR*, abs/2103.15916, 2021. 1, 2

[22] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *CoRR*, abs/2004.12943, 2020. 1

[23] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. *CoRR*, abs/2012.06440, 2020. 2, 4

[24] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati. Deep latent space learning for cross-modal mapping of audio and visual signals. In *DICTA 2019,*, pages 1–7. IEEE, 2019. 2

[25] Phuc Xuan Nguyen, Deva Ramanan, and Charless C. Fowlkes. Weakly-supervised action localization with background modeling. In *2019 IEEE/CVF ICCV Seoul, Korea (South), October 27 - November 2, 2019*, pages 5501–5510. IEEE, 2019. 2, 4

[26] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018*, volume 11210 of *Lecture Notes in Computer Science*, pages 639–658. Springer, 2018. 2

[27] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: weakly-supervised temporal activity localization and classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018*, volume 11208 of *Lecture Notes in Computer Science*, pages 588–607. Springer, 2018. 5

[28] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020,*, pages 4372–4376. IEEE, 2020. 6

[29] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *WACV 2020*, pages 2959–2968. IEEE, 2020. 6

[30] Rui Su, Dong Xu, Lu Sheng, and Wanli Ouyang. Pcg-tal: Progressive cross-granularity cooperation for temporal action localization. *IEEE Transactions on Image Processing*, 30:2103–2113, 2021. 1

[31] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020*, volume 12348 of *Lecture Notes in Computer Science*, pages 436–454. Springer, 2020. 3

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Computer Vision - ECCV 2018*, volume 11206 of *Lecture Notes in Computer Science*, pages 252–268. Springer, 2018. 1, 2, 4, 5, 6

[33] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR 2018*, pages 7794–7803. IEEE Computer Society, 2018. 4

[34] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV 2019*, pages 6291–6299. IEEE, 2019. 1, 2, 5, 6

[35] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *MM '20*, pages 3893–3901. ACM, 2020. 2, 4, 5, 6, 7

[36] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI 2020, New York*, pages 279–286. AAAI Press, 2020. 1, 2, 6

[37] Jingran Zhang, Xing Xu, Fumin Shen, Huimin Lu, Xin Liu, and Heng Tao Shen. Enhancing audio-visual association with self-supervised curriculum learning. In *AAAI 2021*, pages 3351–3359. AAAI Press, 2021. 1, 2

[38] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. *CoRR*, abs/2104.00239, 2021. 1, 3, 6, 7