Audio-Visual Event Localization in Unconstrained Videos

Yapeng Tian[1], Jing Shi[1], Bochen Li[2], Zhiyao Duan[2], and Chenliang Xu[1]

[1]Department of Computer Science, University of Rochester
[2]Department of Electrical and Computer Engineering, University of Rochester

Abstract. In this paper, we introduce a novel problem of audio-visual event localization in unconstrained videos. We define *an audio-visual event* as an event that is both visible and audible in a video segment. We collect an *Audio-Visual Event* (AVE) dataset to systemically investigate three temporal localization tasks: supervised and weakly-supervised audio-visual event localization, and cross-modality localization. We develop an audio-guided visual attention mechanism to explore audio-visual correlations, propose a dual multimodal residual network (DMRN) to fuse information over the two modalities, and introduce an audio-visual distance learning network to handle the cross-modality localization. Our experiments support the following findings: joint modeling of auditory and visual modalities outperforms independent modeling, the learned attention can capture semantics of sounding objects, temporal alignment is important for audio-visual fusion, the proposed DMRN is effective in fusing audio-visual features, and strong correlations between the two modalities enable cross-modality localization.

Keywords: audio-visual event, temporal localization, attention, fusion

1 Introduction

Studies in neurobiology suggest that the perceptual benefits of integrating visual and auditory information are extensive [1]. For computational models, they reflect in lip reading [2,3], where correlations between speech and lip movements provide a strong cue for linguistic understanding; in music performance [4], where vibrato articulations and hand motions enable the association between sound tracks and the performers; and in sound synthesis [5], where physical interactions with different types of material give rise to plausible sound patterns. Albeit these advances, these models are limited in their constrained domains.

Indeed, our community has begun to explore marrying computer vision with audition *in-the-wild* for learning a *good* representation [6,7,8]. For example, a sound network is learned in [6] by a visual teacher network with a large amount of unlabeled videos, which shows better performance than learning in a single modality. However, they have all assumed that the audio and visual contents in a video are matched (which is often not the case as we will show) and they are yet to explore whether the joint audio-visual representations can facilitate understanding unconstrained videos.
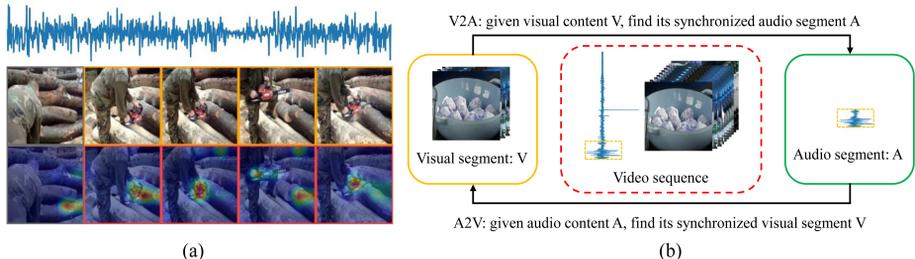
Fig. 1: (a) illustrates audio-visual event localization. The first two rows show a 5s video sequence with both audio and visual tracks for an audio-visual event *chainsaw* (event is temporally labeled in yellow boxes). The third row shows our localization results (in red boxes) and the generated audio-guided visual attention maps. (The first frame does not contain the chainsaw event, hence the attention focuses on background regions.) (b) illustrates cross-modality localization for V2A and A2V

In this paper, we study a family of audio-visual event temporal localization tasks (see Fig. 1) as a proxy to the broader audio-visual scene understanding problem for unconstrained videos. We pose and seek to answer the following questions: (Q1) Does inference jointly over auditory and visual modalities outperform inference over them independently? (Q2) How does the result vary under noisy training conditions? (Q3) How does knowing one modality help model the other modality? (Q4) How do we best fuse information over both modalities? (Q5) Can we locate the content in one modality given its observation in the other modality? Notice that the individual questions might be studied in the literature, but we are not aware of any work that conducts a systematic study to answer these collective questions as a whole.

In particular, we define an *audio-visual event* as an event that is both visible and audible in a video segment, and we establish three tasks to explore aforementioned research questions: 1) supervised audio-visual event localization, 2) weakly-supervised audio-visual event localization, and 3) event-agnostic cross-modality localization. The first two tasks aim to predict which temporal segment of an input video has an audio-visual event and what category the event belongs to. The weakly-supervised setting assumes that we have no access to the temporal event boundary but an event tag at video-level for training. Q1-Q4 will be explored within these two tasks. In the third task, we aim to locate the corresponding visual sound source temporally within a video from a given sound segment and vice versa, which will answer Q5.

We propose both baselines and novel algorithms to solve the above three tasks. For the first two tasks, we start with a baseline model treating them as a sequence labeling problem. We utilize CNN [9] to encode audio and visual inputs, adapt LSTM [10] to capture temporal dependencies, and apply Fully Connected (FC) network to make the final predictions. Upon this baseline model, we introduce an audio-guided visual attention mechanism to verify whether audio can help attend visual features; it also implies spatial locations for sounding

objects as a side output. Furthermore, we investigate several audio-visual feature fusion methods and propose a novel dual multimodal residual fusion network that achieves the best fusion results. For weakly-supervised learning, we formulate it as a Multiple Instance Learning (MIL) [11] task, and modify our network structure via adding a MIL pooling layer to handle the problem. To address the harder cross-modality localization task, we propose an audio-visual distance learning network that measures the relativeness of any given pair of audio and visual content. It projects audio and visual features into subspaces with the same dimension. Contrastive loss [12] is introduced to learn the network.

Observing that there is no publicly available dataset directly suitable for our tasks, we collect a large video dataset that consists of 4143 10-second videos with both audio and video tracks for 28 audio-visual events and annotate their temporal boundaries. Videos in our dataset are originated from YouTube, thus they are unconstrained. Our extensive experiments support the following findings: modeling jointly over auditory and visual modalities outperforms modeling independently over them, audio-visual event localization in a noisy condition can still achieve promising results, the audio-guided visual attention can well capture semantic regions covering sounding objects and can even distinguish audio-visual unrelated videos, temporal alignment is important for audio-visual fusion, the proposed dual multimodal residual network is effective in addressing the fusion task, and strong correlations between the two modalities enable cross-modality localization. These findings have paved a way for our community to solve harder, high-level understanding problems in the future, such as video captioning [13] and movieQA [14], where the auditory modality plays an important role in understanding video but lacks effective modeling.

Our work makes the following contributions: (1) a family of three audio-visual event localization tasks; (2) an audio-guided visual attention model to adaptively explore the audio-visual correlations; (3) a novel dual multimodal residual network to fuse audio-visual features; (4) an effective audio-visual distance learning network to address cross-modality localization; (5) a large audio-visual event dataset containing more than 4K unconstrained and annotated videos, which to the best of our knowledge, is the largest dataset for sound event detection. Dataset, code, and supplementary material are available on our webpage: https://sites.google.com/view/audiovisualresearch.

2   Related Work

In this section, we first describe how our work differs from closely-related topics: sound event detection, temporal action localization and multimodal machine learning, then discuss relations to recent works in modeling vision-and-sound.

Sound event detection considered in the audio signal processing community aims to detect and temporally locate sound events in an acoustic scene. Approaches based on Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), feed-forward Deep Neural Networks (DNN), and Bidirectional Long Short-Term Memory (BLSTM) [15] are developed in [16,17,18,19]. These meth-

ods focus on audio signals, and visual signals have not been explored. Corresponding datasets, e.g., TUT [20], for sound event detection only contain sound tracks, and are not suitable for audio-visual scene understanding.

Temporal action localization aims to detect and locate actions in videos. Most works cast it as a classification problem and utilize a temporal sliding window approach, where each window is considered as an action candidate subject to classification [21]. Escorcia *et al.* [22] present a deep action proposal network that is effective in generating temporal action proposals for long videos and can speed up temporal action localization. Recently, Shou *et al.* [23] propose an end-to-end Segment-based 3D CNN method (S-CNN), Zhao *et al.* [24] present a structured segment network (SSN), and Lea *et al.* [25] develop an Encoder-Decoder Temporal Convolutional Network (ED-TCN) to hierarchically model actions. Different from these works, an audio-visual event in our consideration may contain multiple actions or motionless sounding objects, and we model over both audio and visual domains. Nevertheless, we extend the ED-TCN and SSN methods to address our supervised audio-visual event localization task and compare them in Sec. 6.3.

Multimodal machine learning aims to learn joint representations over multiple input modalities, e.g., speech and video, image and text. Feature fusion is one of the most important part for multimodal learning [26], and many different fusion models have been developed, such as statistical models [27], Multiple Kernel Learning (MKL) [28,29], Graphical models [30,31]. Although some mutimodal deep networks have been studied in [31,32,33,34,35,36,37], which mainly focus on joint audio-visual representation learning based on Autoencoder or deep Boltzmann machines [32], we are interested in investigating the best models to fuse learned audio and visual features for localization purpose.

Recently, some inspiring works are developed for modeling vision-and-sound [8,6,7,5,38]. Aytar *et al.* [6] use a visual teacher network to learn powerful sound representations from unlabeled videos. Owens *et al.* [7] leverage ambient sounds as supervision to learn visual representations. Arandjelovic and Zisserman [8] learn both visual and audio representations in an unsupervised manner through an audio-visual correspondence task, and in [39], they further locate sound source spatially in an image based on an extended correspondence network. Aside from works in representation learning, audio-visual cross-modal synthesis is studied in [7,40,41], and associations between natural image scenes and accompanying free-form spoken audio captions are explored in [38]. Concurrently, some interesting and related works on sound source separation, localization and audio-visual representation learning are explored in [42,43,44,45,46]. Unlike the previous works, in this paper, we systematically investigate audio-visual event localization tasks.

3   Dataset and Problems

AVE: The *Audio-Visual Event* Dataset To the best of our knowledge, there is no publicly available dataset directly suitable for our purpose. Therefore, we introduce the *Audio-Visual Event* (AVE) dataset , a subset of AudioSet [47], that

Fig. 2: The AVE dataset. Some examples in the dataset are shown. The distribution of videos in different categories and the distribution of event lengths are illustrated

contains 4143 videos covering 28 event categories and videos in AVE are temporally labeled with audio-visual event boundaries. Each video contains at least one 2s long *audio-visual event*. The dataset covers a wide range of audio-visual events (*e.g.*, man speaking, woman speaking, dog barking, playing guitar, and frying food *etc.*) from different domains, e.g., human activities, animal activities, music performances, and vehicle sounds. We provide examples from different categories and show the statistics in Fig. 2. Each event category contains a minimum of 60 videos and a maximum of 188 videos, and 66.4% videos in the AVE contain audio-visual events that span over the full 10 seconds. Next, we introduce three different tasks based on the AVE to explore the interactions between auditory and visual modalities.

Fully and Weakly-Supervised Event Localization The goal of event localization is to predict the event label for each video segment, which contains both audio and visual tracks, for an input video sequence. Concretely, for a video sequence, we split it into T non-overlapping segments $\{V_t, A_t\}_{t=1}^{T}$, where each segment is 1s long (since our event boundary is labeled at second-level), and V_t and A_t denote the visual content and its corresponding audio counterpart in a video segment, respectively. Let $\boldsymbol{y}_t = \{y_t^k | y_t^k \in \{0,1\}, k = 1, ..., C, \sum_{k=1}^{C} y_t^k = 1\}$ be the event label for that video segment. Here, C is the total number of AVE events plus one background label.

For the supervised event localization task, the event label $\boldsymbol{y}_t$ of each visual segment V_t or audio segment A_t is known during training. We are interested in event localization in audio space alone, visual space alone and the joint audio-visual space. This task explores whether or not audio and visual information can help each other improve event localization. Different than the supervised setting, in the weakly-supervised manner we have only access to a video-level event tag, and we still aim to predict segment-level labels during testing. The weakly-supervised task allows us to alleviate the reliance on well-annotated data for modelings of audio, visual and audio-visual.

Cross-Modality Localization In the cross-modality localization task, given a segment of one modality (auditory/visual), we would like to find the position of its synchronized content in the other modality (visual/auditory). Concretely, for visual localization from audio (A2V), given a l-second audio segment $\hat{A}$ from $\{A_t\}_{t=1}^{T}$, where $l < T$, we want to find its synchronized l-second visual segment
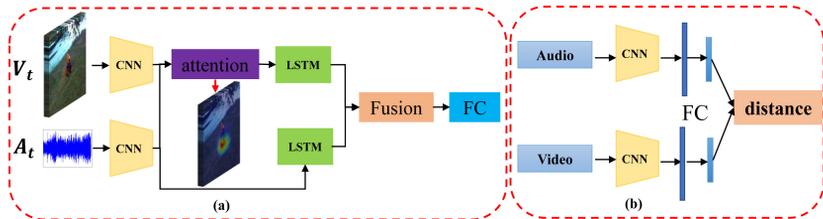
Fig. 3: (a) Audio-visual event localization framework with audio-guided visual attention and multimodal fusion. One timestep is illustrated, and note that the fusion network and FC are shared for all timesteps. (b) Audio-visual distance learning network

within $\{V_t\}_{t=1}^T$. Similarly, for audio localization from visual content (V2A), given a l-second video segment $\hat{V}$ from $\{V_t\}_{t=1}^T$, we would like to find its l-second audio segment within $\{A_t\}_{t=1}^T$. This task is conducted in the event-agnostic setting such that the models developed for this task are expected to work for general videos where the event labels are not available. For evaluation, we only use short-event videos, in where the lengths of audio-visual event are all shorter than 10s.

4   Methods for Audio-Visual Event Localization

First, we present the overall framework that treats the audio-visual event localization as a sequence labeling problem in Sec. 4.1. Upon this framework, we propose our audio-guided visual attention in Sec. 4.2 and a novel dual multimodal residual fusion network in Sec. 4.3. Finally, we extend this framework to work in weakly-supervised setting in Sec. 4.4.

4.1   Audio-Visual Event Localization Network

Our network mainly consists of five modules: feature extraction, audio-guided visual attention, temporal modeling, multimodal fusion and temporal labeling (see Fig. 3(a)). The feature extraction module utilizes pre-trained CNNs to extract visual features $v_t = [v_t^1, ..., v_t^k] \in \mathbb{R}^{d_v \times k}$ and audio features $a_t \in \mathbb{R}^{d_a}$ from each V_t and A_t, respectively. Here, d_v denotes the number of CNN visual feature maps, k is the vectorized spatial dimension of each feature map, and d_a denotes the dimension of audio features. We use an audio-guided visual attention model to generate a context vector $v_t^{att} \in \mathbb{R}^{d_v}$ (see details in Sec. 4.2). Two separate LSTMs take v_t^{att} and a_t as inputs to model temporal dependencies in the two modalities respectively. For an input feature vector F_t at time step t, the LSTM updates a hidden state vector h_t and a memory cell state vector c_t:

$$h_t, c_t = \mathrm{LSTM}(F_t, h_{t-1}, c_{t-1}) \ , \tag{1}$$

where F_t refers to v_t^{att} or a_t in our model. For evaluating the performance of the proposed attention mechanism, we compare to models that do not use attention;

we directly feed global average pooling visual features and audio features into LSTMs as baselines. To better incorporate the two modalities, we introduce a multimodal fusion network (see details in Sec. 4.3). The audio-visual representation h_t^* is learned by a multimodal fusion network with audio and visual hidden state output vectors h_t^v and h_t^a as inputs. This joint audio-visual representation is used to output event category for each video segment. For this, we use a shared FC layer with the Softmax activation function to predict probability distribution over C event categories for the input segment and the whole network can be trained with a multi-class cross-entropy loss.

4.2 Audio-Guided Visual Attention

Psychophysical and physiological evidence shows that sound is not only informative about its source but also its location [48]. Based on this, Hershey and Movellan [49] introduce an exploratory work on localizing sound sources utilizing audio-visual synchrony. It shows that the strong correlations between the two modalities can be used to find image regions that are highly correlated to the audio signal. Recently, [7,39] show that sound indicates object properties even in unconstrained images or videos. These works inspire us to use audio signal as a means of guidance for visual modeling.

Given that attention mechanism has shown superior performance in many applications such as neural machine translation [50] and image captioning [51,52], we use it to implement our audio-guided visual attention (see Fig. 3(a) and Fig. 4(a)). The attention network will adaptively learn which visual regions in each segment of a video to look for the corresponding sounding object or activity.

Fig. 4: (a) Audio-guided visual attention mechanism. (b) Dual multimodal residual network for audio-visual feature fusion

Concretely, we define the attention function f_{att} and it can be adaptively learned from the visual feature map v_t and audio feature vector a_t. At each time step t, the visual context vector v_t^{att} is computed by:

$$v_t^{att} = f_{att}(a_t, v_t) = \sum_{i=1}^{k} w_t^i v_t^i \ , \tag{2}$$

where w_t is an attention weight vector corresponding to the probability distribution over k visual regions that are attended by its audio counterpart. The attention weights can be computed based on MLP with a Softmax activation function:

$$w_t = Softmax(x_t) \ , \tag{3}$$

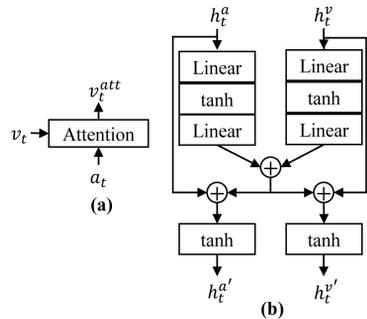$$x_t = W_f \sigma(W_v U_v(v_t) + (W_a U_a(a_t)) \mathbb{1}^T) \ , \tag{4}$$

where U_v and U_a, implemented by a dense layer with nonlinearity, are two transformation functions that project audio and visual features to the same dimension d, $W_v \in \mathbb{R}^{k \times d}$, $W_a \in \mathbb{R}^{k \times d}$, $W_f \in \mathbb{R}^{1 \times k}$ are parameters, the entries in $\mathbb{1} \in \mathbb{R}^k$ are all 1, $\sigma(\cdot)$ is the hyperbolic tangent function, and $w_t \in \mathbb{R}^k$ is the computed attention map. The attention map visualization results show that the audio-guided attention mechanism can adaptively capture the location information of sound source (see Fig. 5), and it can also improve temporal localization accuracy (see Tab. 1).

4.3   Audio-Visual Feature Fusion

Our fusion method is designed based on the philosophy in [32], which processes multiple features separately and then learns a joint representation using a middle layer. To combine features coming from visual and audio modalities, inspired by the Mutimodal Residual Network (MRN) in [53] (which works for text-and-image), we introduce a Dual Multimodal Residual Network (DMRN). The MRN adopts a textual residual branch and feeds transformed visual features into different textual residual blocks, where only textual features are updated. In contrary, the proposed DMRN shown in Fig. 4(b) updates both audio and visual features simultaneously.

Given audio and visual features h_t^a and h_t^v from LSTMs, the DMRN will compute the updated audio and visual features:

$$h_t^{a'} = \sigma(h_t^a + f(h_t^a, h_t^v)) \ , \tag{5}$$

$$h_t^{v'} = \sigma(h_t^v + f(h_t^a, h_t^v)) \ , \tag{6}$$

where $f(\cdot)$ is an additive fusion function, and the average of $h_t^{a'}$ and $h_t^{v'}$ is used as the joint representation h_t^* for labeling the video segment. Here, the update strategy in DMRN can both preserve useful information in the original modality and add complimentary information from the other modality. Simply, we can stack multiple residual blocks to learn a deep fusion network with updated $h_t^{a'}$ and $h_t^{v'}$ as inputs of new residual blocks. However, we empirically find that it does not improve performance by stacking many blocks for both MRN and DMRN. We argue that the network becomes harder to train with increasing parameters and one block is enough to handle this simple fusion task well.

We would like to underline the importance of fusing audio-visual features after LSTMs for our task. We empirically find that late fusion (fusion after temporal modeling) is much better than early fusion (fusion before temporal modeling). We suspect that the auditory and visual modalities are not temporally aligned. Temporal modeling by LSTMs can implicitly learn certain alignments which can help make better audio-visual fusion. The empirical evidences will be shown in Tab. 2.

4.4   Weakly-Supervised Event Localization

To address the weakly-supervised event localization, we formulate it as a MIL problem and extend our framework to handle noisy training condition. Since only video-level labels are available, we infer label of each audio-visual segment pair in the training phase, and aggregate these individual predictions into a video-level prediction by MIL pooling as in [54]:

$$\hat{m} = g(m_1, m_2, ..., m_T) = \frac{1}{T}\sum_{t=1}^{T} m_t \ , \tag{7}$$

where $m_1, ..., m_T$ are predictions from the last FC layer of our audio-visual event localization network, and $g(\cdot)$ averages over all predictions. The probability distribution of event category for the video sequence can be computed using $\hat{m}$ over the Softmax. During testing, we can predict the event category for each segment according to computed m_t.

5   Method for Cross-Modality Localization

To address the cross-modality localization problem, we propose an audio-visual distance learning network (AVDLN) as illustrated in Fig. 3(b); we notice similar networks are studied in concurrent works [39,55]. Our network can measure the distance $D_\theta(V_i, A_i)$ for a given pair of V_i and A_i. At test time, for visual localization from audio (A2V), we use a sliding window method and optimize the following objective:

$$t^* = \underset{t}{\operatorname{argmin}} \sum_{s=1}^{l} D_\theta(V_{s+t-1}, \hat{A}_s) \ , \tag{8}$$

where $t^* \in \{1, ..., T-l+1\}$ denotes the start time when visual and audio content synchronize, T is the total length of a testing video sequence, and l is the length of the audio query $\hat{A}$. This objective function computes an optimal matching by minimizing the cumulative distance between the audio segments and the visual segments. Therefore, $\{V_i\}_{i=t^*}^{t^*+l-1}$ is the matched visual content. Similarly, we can define audio localization from visual content (V2A); we omit it here for a concise writing. Next, we describe the network used to implement the matching function.

Let $\{V_i, A_i\}_{i=1}^{N}$ be N training samples and $\{y_i\}_{i=1}^{N}$ be their labels, where V_i and A_i are a pair of 1s visual and audio segments, $y_i \in \{0, 1\}$. Here, $y_i = 1$ means that V_i and A_i are synchronized. The AVDLN will learn to measure distances between these pairs. The network encodes them using pre-trained CNNs, and then performs dimensionality reduction for encoded audio and visual representations using two different two-layer FC networks. The outputs of final FC layers are $\{R_i^v, R_i^a\}_{i=1}^{N}$. The distance between V_i and A_i is measured by the Euclidean distance between R_i^v and R_i^a:

$$D_\theta(V_i, A_i) = ||R_i^v - R_i^a||_2 \ . \tag{9}$$

To optimize the parameters θ of the distance metric D_θ, we introduce the contrastive loss proposed by Hadsell *et al.* [12]. The contrastive loss function is:

$$L_C = y_i D_\theta^2(V_i, A_i) + (1 - y_i)(\max(0, th - D_\theta(V_i, A_i)))^2, \qquad (10)$$

where $th > 0$ is a margin. If a dissimilar pair's distance is less than th, the loss will make the distance D_θ bigger; if their distance is bigger than the margin, it will not contribute to the loss.

6   Experiments

First, we introduce the used visual and audio representations in Sec. 6.1. Then, we describe the compared baseline models and evaluation metrics in Sec 6.2. Finally, we show and analyze experimental results of different models in Sec. 6.3.

6.1   Visual and Audio Representations

It has been suggested that CNN features learned from a large-scale dataset (*e.g.* ImageNet [56], AudioSet [47]) are highly generic and powerful for other vision or audition tasks. So, we adopt pre-trained CNN models to extract features for visual segments and their corresponding audio segments.

For each 1s visual segment, we extract *pool*5 feature maps from sampled 16 RGB video frames by VGG-19 network [57], which is pre-trained on ImageNet, and then utilize global average pooling [58] over the 16 frames to generate one $512 \times 7 \times 7$-D feature map. We also explore the temporal visual features extracted by C3D [59], which is capable of learning spatio-temporal visual features. But we do not observe significant improvements when combining C3D features. We extract a 128-D audio representation for each 1s audio segment via a VGG-like network [60] pre-trained on AudioSet.

6.2   Baselines and Evaluation Metrics

To validate the effectiveness of the joint audio-visual modeling, we use single-modality models as baselines, which only use audio-alone or visual-alone features and share the same structure with our audio-visual models. To evaluate the audio-guided visual attention, we compare our V-att and A+V-att models with V and A+V models in fully and weakly supervised settings. Here, V-att models adopt audio-guided visual attention to pool visual feature maps, and the other V models use global average pooling to compute visual feature vectors. We visualize generated attention maps for subjective evaluation. To further demonstrate the effectiveness of the proposed networks, we also compare them with a state-of-the-art temporal labeling network: ED-TCN [25] and proposal-based SSN [24].

We compare our fusion method: DMRN with several network-based multi-modal fusion methods: Additive, Maxpooling (MP), Gated, Multimodal Bilinear (MB), and Gated Multimodal Bilinear (GMB) in [37], Gated Multimodal Unit
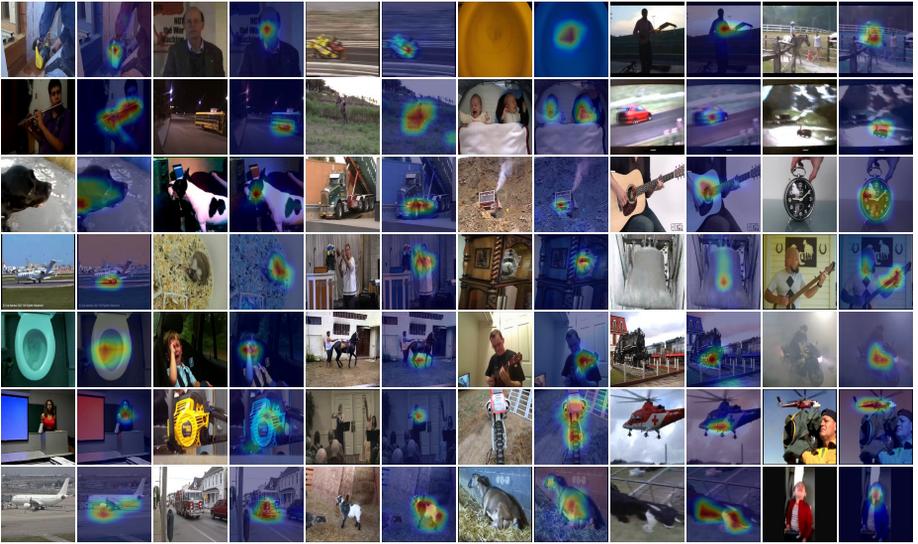
Fig. 5: Qualitative visualization of audio-guided visual attention. The semantic regions containing many different sound sources, such as barking dog, crying boy/babies, speaking woman, horning bus, guitar *etc*, can be adaptively captured by our attention model
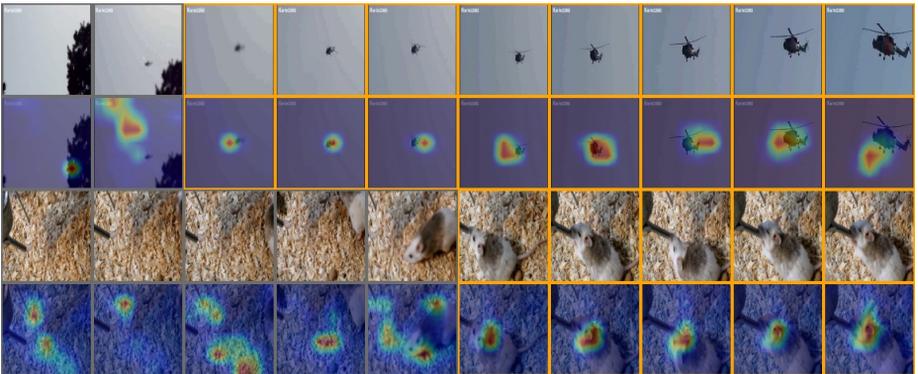

Fig. 6: Visualization of visual attention maps on two **challenging** examples. The first and third rows are 10 video frames uniformly extracted from two 10s videos, and the second and fourth rows are generated attention maps. The **yellow** box (groundtruth label) denotes that the frame contain audio-visual event in which sounding object is visible and sound is audible. If there is no audio-visual event in a frame, random background regions will be attended (5th frame of the second example); otherwise, the attention will focus on sounding sources

(GMU) in [61], Concatenation (Concat), and MRN [53]. Three different fusion strategies: early, late and decision fusions are explored. Here, early fusion methods directly fuse audio features from pre-trained CNNs and attended visual

Table 1: Event localization prediction accuracy (%) on AVE dataset. A, V, V-att, A+V, A+V-att denote that these models use audio, visual, attended visual, audio-visual and attended audio-visual features, respectively. W-models are trained in a weakly-supervised manner. Note that audio-visual models all fuse features by concatenating the outputs of LSTMs

Models	A	V	V-att	A+V	A+V-att	W-A	W-V	W-V-att	W-A+V	W-A+V-att
Accuracy	59.5	55.3	58.6	71.4	**72.7**	53.4	52.9	55.6	63.7	66.7

features; late fusion methods fuse audio and visual features from outputs of two LSTMs; and decision fusion methods fuse the two modalities before Softmax layer. To further enhance the performance of DMRN, we also introduce a variant model of DMRN called dual multimodal residual fusion ensemble (DMRFE) method, which feeds audio and visual features into two separated blocks and then use average ensemble to combine the two predicted probabilities.

For supervised and weakly-supervised event localization, we use overall accuracy as an evaluation metric. For cross-modality localization, e.g., V2A and A2V, if a matched audio/visual segment is exactly the same as its groundtruth, we regard it as a good matching; otherwise, it will be a bad matching. We compute the percentage of good matchings over all testing samples as prediction accuracy to evaluate the performance of cross-modality localization. To validate the effectiveness of the proposed model, we also compare it with deep canonical correlation analysis (DCCA) method [62].

6.3   Experimental Comparisons

Table 1 compares different variations of our proposed models on supervised and weakly-supervised audio-visual event localization tasks. Table 2 shows event localization performance of different fusion methods. Figures 5 and 6 illustrate generated audio-guided visual attention maps.

To benchmark our models with state-of-the-art temporal action localization methods, we extend the SSN [24] and ED-TCN [25] to address the supervised audio-visual event localization, and train them on AVE. The SSN and ED-TCN achieve 26.7% and 46.9% overall accuracy, respectively. For comparison, our V model with the same features achieves 55.3%.

Audio and Visual. From Tab. 1, we observe that A outperforms V and W-A is also better than W-V. It demonstrates that audio features are more powerful to address audio-visual event localization task on the AVE dataset. However, when we look at each individual event, using audio is not always better than using visual. We observe that V is better than A for some events (*e.g.* car, motocycle, train, bus). Actually, most of these events are outdoor. Audios in these videos can be very noisy: several different sounds may be mixed together (*e.g.* people cheers with a racing car), and may have very low intensity (*e.g.* horse sound from far distance). For these conditions, visual information will give

Table 2: Event localization prediction accuracy (%) of different feature fusion methods on AVE dataset. These methods all use same audio and visual features as inputs. Top-2 results in each line are highlighted

Methods	Additive	MP	Gated	MB	GMU	GMB	Concat	MRN	DMRN	DMRFE
Early Fusion	59.9	67.9	67.9	69.2	**70.5**	**70.2**	61.0	69.8	68.0	-
Late Fusion	71.3	71.4	70.5	70.5	71.6	71.0	72.7	70.8	**73.1**	**73.3**
Decision Fusion	**70.5**	64.5	65.2	64.6	67.6	67.3	69.7	63.8	**70.4**	-

us more discriminative and accurate information to understand events in videos. A is much better than V for some events (*e.g.* dog, man and woman speaking, baby crying). Sounds will provide clear cues for us to recognize these events. For example, if we hear barking sound, we know that there may be a dog. We also observe that A+V is better than both A and V, and W-A+V is better than W-A and W-V. From the above results and analysis, we can conclude that auditory and visual modalities will provide complementary information for us to understand events in videos. The results also demonstrate that our AVE dataset is suitable for studying audio-visual scene understanding tasks.

Audio-Guided Visual Attention. The quantitative results (see Tab. 1) show that V-att is much better than V (a 3.3% absolute improvement) and A+V-att outperforms A+V by 1.3%. We show qualitative results of our attention method in Fig. 5. We observe that a range of semantic regions in many different categories and examples can be attended by sound, which validates that our attention network can learn which visual regions to look at for sounding objects. An interesting observation is that the audio-guided visual attention tends to focus on sounding regions, such as man's mouth, head of crying boy *etc*, rather than whole objects in some examples. Figure 6 illustrates two challenging cases. For the first example, the sounding helicopter is quite small in the first several frames but our attention model can still capture its locations. For the second example, the first five frames do not contain an audio-visual event; in this case, attentions are spread on different background regions. When the rat appears in the 5th frame but is not making any sound, the attention does not focus on the rat. When the rat sound becomes audible, the attention focuses on the sounding rat. This observation validates that the audio-guided attention mechanism is helpful to distinguish audio-visual unrelated videos, and is not just to capture a saliency map with objects.

Audio-Visual Fusion. Table 2 shows audio-visual event localization prediction accuracy of different multimodal feature fusion methods on AVE dataset. Our DMRN model in the late fusion setting can achieve better performance than all compared methods, and our DMRFE model can further improve performance. We also observe that late fusion is better than early fusion and decision fusion. The superiority of late fusion over early fusion demonstrates that temporal modeling before audio-visual fusion is useful. We know that auditory and visual

modalities are not completely aligned, and temporal modeling can implicitly learn certain alignments between the two modalities, which is helpful for the audio-visual feature fusion task. The decision fusion can be regard as a type of late fusion but using lower dimension (same as the category number) features. The late fusion outperforms the decision fusion, which validates that processing multiple features separately and then learning joint representation using a middle layer rather than the bottom layer is an efficient fusion way.

Full and Weak Supervision. Obviously, supervised models are better than weakly supervised ones, but quantitative comparisons show that weakly-supervised approaches achieve promising event localization performance, which demonstrates the effectiveness of the MIL frameworks, and validates that the audio-visual event localization task can be addressed even in a noisy condition.

Cross-Modality Localization. Table 3 reports the prediction accuracy of our method and DCCA [62] on cross-modality localization task. Our AVDL outperforms DCCA over a large margin both on A2V and V2A tasks. Even using the strict evaluation metric (which counts only the exact matches), our models on both subtasks: A2V and V2A, show promising results, which further demonstrates that there are strong correlations between audio and visual modalities, and it is possible to address the cross-modality localization for unconstrained videos.

Table 3: Accuracy on cross-modality localization. A2V: visual localization from audio segment query; V2A: audio localization from visual segment query

Models	AVDLN	DCCA
A2V	**44.8**	34.8
V2A	**35.6**	34.1

7   Conclusion

In this work, we study a suit of five research questions in the context of three audio-visual event localization tasks. We propose both baselines and novel algorithms to address each of the three tasks. Our systematic study well supports our findings: modeling jointly over auditory and visual modalities outperforms independent modeling, audio-visual event localization in a noisy condition is still tractable, the audio-guided visual attention is able to capture semantic regions of sound sources and can even distinguish audio-visual unrelated videos, temporal alignments are important for audio-visual feature fusion, the proposed dual residual network is capable of audio-visual fusion, and strong correlations existing between the two modalities enable cross-modality localization.

References

1. Bulkin, D.A., Groh, J.M.: Seeing sounds: visual and auditory interactions in the brain. Current opinion in neurobiology **16**(4) (2006) 415–419
2. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. CoRR **abs/1611.01599** (2016)
3. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. (2017)
4. Li, B., Xu, C., Duan, Z.: Audio-visual source association for string ensembles through multi-modal vibrato analysis. In: Sound and Music Computing. (2017)
5. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2405–2413
6. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. (2016) 892–900
7. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European Conference on Computer Vision. (2016) 801–816
8. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. (2017)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
11. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Advances in neural information processing systems. (1998) 570–576
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proc. CVPR. (2006)
13. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016)
14. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016)
15. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11) (1997) 2673–2681
16. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing **2013**(1) (2013) 1
17. Mesaros, A., Heittola, T., Virtanen, T.: Tut database for acoustic scene classification and sound event detection. In: Signal Processing Conference (EUSIPCO), 2016 24th European, IEEE (2016) 1128–1132
18. Cakir, E., Heittola, T., Huttunen, H., Virtanen, T.: Polyphonic sound event detection using multi label deep neural networks. In: Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE (2015) 1–7
19. Parascandolo, G., Huttunen, H., Virtanen, T.: Recurrent neural networks for polyphonic sound event detection in real life recordings. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE (2016) 6440–6444

20. Mesaros, A., Heittola, T., Virtanen, T.: Tut database for acoustic scene classification and sound event detection. In: EUSIPCO. (Aug 2016)
21. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: Proceedings of the IEEE international conference on computer vision. (2013) 1817–1824
22. Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B.: Daps: Deep action proposals for action understanding. In: European Conference on Computer Vision, Springer (2016) 768–784
23. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
24. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. ICCV (2017)
25. Lea *et al.*, C.: Temporal convolutional networks for action segmentation and detection. In: CVPR. (2017)
26. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
27. Fisher III, J.W., Darrell, T., Freeman, W.T., Viola, P.A.: Learning joint statistical models for audio-visual fusion and segregation. In: Advances in neural information processing systems. (2001) 772–778
28. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of machine learning research **12**(Jul) (2011) 2211–2268
29. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing. (2015) 2539–2544
30. Gurban, M., Thiran, J.P., Drugman, T., Dutoit, T.: Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In: Proceedings of the 10th international conference on Multimodal interfaces, ACM (2008) 237–240
31. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. (2011)
32. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems. (2012) 2222–2230
33. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop. (2012)
34. Mroueh, Y., Marcheret, E., Goel, V.: Deep multimodal learning for audio-visual speech recognition. In: ICASSP, IEEE (2015)
35. Hu, D., Li, X., et al.: Temporal multimodal learning in audiovisual speech recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3574–3582
36. Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E.A., Luo, J.: Deep multimodal representation learning from temporal data. In: CVPR. (July 2017)
37. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient large-scale multi-modal classification. In: AAAI. (2018)
38. Harwath, D., Torralba, A., Glass, J.: Unsupervised learning of spoken language with visual context. In: Advances in Neural Information Processing Systems. (2016) 1858–1866

39. Arandjelović, R., Zisserman, A.:   Objects that sound.   arXiv preprint arXiv:1712.06651 (2017)
40. Yipin Zhou, Zhaowen Wang, C.F.T.B., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: CVPR. (2018)
41. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia. (2017)
42. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: CVPR. (2018) 4358–4366
43. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.:   Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation.   arXiv preprint arXiv:1804.03619 (2018)
44. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV. (2018)
45. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV. (2018)
46. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV. (2018)
47. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. (2017)
48. Gaver, W.W.: What in the world do we hear?: An ecological approach to auditory event perception. Ecological psychology **5**(1) (1993) 1–29
49. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: Advances in neural information processing systems. (2000) 813–819
50. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR. (2015)
51. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICLR. (2015)
52. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR. (2017)
53. Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T.: Multimodal residual learning for visual qa. In: NIPS. (2016) 361–369
54. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proc. CVPR. (2015) 3460–3469
55. Surís, D., Duarte, A., Salvador, A., Torres, J., Giró-i Nieto, X.: Cross-modal embeddings for video and audio retrieval. arXiv preprint arXiv:1801.02200 (2018)
56. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 211–252
57. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
58. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
59. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2015) 4489–4497
60. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP. (2017) 131–135

61. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. In: ICLR workshop. (2017)
62. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning. (2013) 1247–1255