Audio-Visual Event Localization via Recursive Fusion by Joint Co-Attention

Bin Duan[1] Hao Tang[2] Wei Wang[2] Ziliang Zong[3] Guowei Yang[3] Yan Yan[1]
[1]Illinois Institute of Technology, USA
[2]University of Trento, Italy [3]Texas State University, USA
tuffrr5@gmail.com, {hao.tang, wei.wang}@unitn.it, {ziliang, gyang}@txstate.edu, yyan34@iit.edu

Abstract

The major challenge in audio-visual event localization task lies in how to fuse information from multiple modalities effectively. Recent works have shown that the attention mechanism is beneficial to the fusion process. In this paper, we propose a novel joint attention mechanism with multimodal fusion methods for audio-visual event localization. Particularly, we present a concise yet valid architecture that effectively learns representations from multiple modalities in a joint manner. Initially, visual features are combined with auditory features and then turned into joint representations. Next, we make use of the joint representations to attend to visual features and auditory features, respectively. With the help of this joint co-attention, new visual and auditory features are produced, and thus both features can enjoy the mutually improved benefits from each other. It is worth noting that the joint co-attention unit is recursive meaning that it can be performed multiple times for obtaining better joint representations progressively. Extensive experiments on the public AVE dataset have shown that the proposed method achieves significantly better results than the state-of-the-art methods.

1. Introduction

Humans explore the surroundings with their advanced sensory system in daily life, e.g., eyes, ears, and noses. Heterogeneous information from various sensors floods into the human perceptual system, among which sound and vision are two dominant components. In multimodal machine learning, it turns out that the joint learning of audio and visual modalities usually achieves better performance than using single modality for various tasks, e.g., sound localization [1, 12, 23, 46, 24], sound source separation [7, 42, 9, 23, 46, 16, 43, 10, 27] and audio-visual event localization [19, 36, 38].

In this paper, we focus on the audio-visual event localization task. As shown in Fig. 1, an Audio-Visual Event (AVE) is defined in a video sequence that is both audible and visible. The audio-visual event localization task consists of two sub-tasks, one of which is to predict the event label while the other is to predict which segment of the video sequence has an audio-visual event of interest. As in the AVE definition, localizing an AVE must deal with heterogeneous information from both audio and visual modalities. Moreover, recent works [19, 36, 38] show that the performance after fusion outperforms the one that only uses a single modality. Although these approaches present interesting explorations, how to smartly fuse representations from both modalities is still a challenging task.

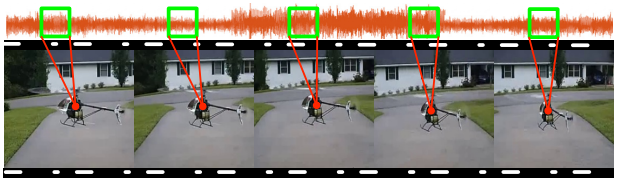

Figure 1. Audio-Visual Event (AVE) is an event both audible and visible. e.g., a person can see a helicopter in the visual sequence (the bottom row) and also hear the helicopter's engine sound in the audio sequence (the top row).

Multimodal fusion provides a global view of multiple representations for a specific phenomenon. To tackle the AVE localization problem, existing methods [19, 36] either fuse cell states out of LSTMs [36], or fuse both hidden states and cell states from LSTMs [19]. Both aforementioned approaches exploit a plain multimodal fusion strategy, where the fusion results might be unstable as it is hard to guarantee good quality of the information used for the fusion, e.g., some noisy information from the background segments may also be included. Therefore, a more robust fusion strategy is needed for better representations. Wu *et al.* [38] introduce a cross-modal matching mechanism that exploits global temporal co-occurrences between two modalities and excludes the noisy background segments from the sequence. Intuitively, having global features to interact with local features would help to localize the event, but it needs additional supervision to manually filter the background segments.

To summarize, existing methods either follow a straightforward multimodal fusion strategy (fuse both features directly), or require extra supervision (exclude background segments). Taking advantage of recursive fusion interactions between multimodal representations, we propose a novel joint co-attention fusion approach that is able to learn more robust representations with less supervision on excluding background segments.

Attention mechanism has been applied to many tasks [45, 34, 8, 41, 6, 3, 21, 33]. For example, recent works in generative adversarial networks [45, 34, 35] utilize a self-attention mechanism that relates different portions of a single image to compute a representation for itself. Besides self-attention, other works in Video Question Answering (VQA) [20, 22] propose a co-attention mechanism, in which the image representation guides the text attention generation and in the meanwhile, the text representation also guides image attention generation. Moreover, both attention mechanisms allow attention-driven, long-range dependency modeling for their corresponding tasks. Motivated by these two attention techniques, we propose a new Joint Co-Attention (JCA) mechanism which develops on the basis of self-attention and co-attention. We utilize the joint representation to generate the attention masks for two uni-modalities while previous methods [20, 22] independently generate attention mask for each other. In our approach, instead of using features from one single modality, each attention mask is generated using features from both modalities and thus it is more informative. As a result, each modality is attended not only by the features from itself (self-attended), but also by the features from the other modality (co-attended).

While the attention mechanism allows multimodal fusion in depth, we further introduce a double fusion mechanism, that can be integrated with attention mechanisms, allowing fusion both in depth and breadth. Existing works [17, 40] exploit the double fusion to integrate representations from different modalities in a hybrid fusion manner, i.e., they fuse features using both early fusion (before feature embedding) and late fusion (after feature embedding). Different from existing methods [19, 36, 17, 40] that fuse representations from multiple modalities simply by averaging, weighting, or concatenation. In this paper, we propose to integrate the double fusion method with our JCA mechanism. First, the audio-guided attention [36] is performed as early fusion. Then, we exploit Bi-LSTM [28] with residual embedding to extract features where we combine features before Bi-LSTM and after Bi-LSTM, leading to global temporal cues. After the Bi-LSTMs, the representations of two modalities are fused using the JCA mechanism as late fusion. Note that the JCA unit is recursive so that the joint co-attention process can be repeated for multiple times.

Overall, our contributions in this paper are summarized as follows:

- We revisit the audio-visual event localization task and tackle the task from a multimodal fusion perspective which targets for better representations.
- We propose a novel joint co-attention mechanism and deploy it in deep audio-visual learning. It learns more robust representations by recursively performing fusions of the representations from two modalities.
- The integration of attention mechanism and double fusion method enables the model to learn long-range dependencies. Extensive experiments show the superiority of our framework.

2. Related Work

Audio-Visual Event Localization aims to identify the event of interest in a video sequence and predict what category the event belongs to. Tian *et al.* [36] first define audio-visual event localization problem aiming to detect event which is both audible and visible. They design an audio-guided attention dual-LSTM network that captures each uni-modal representation and fuses them by concatenation for the final prediction. Lin *et al.* [19] propose a dual-modality sequence-sequence framework that explores the global features of audio and visual modalities. Wu *et al.* [38] introduce a dual attention matching mechanism that conducts cross-matching across modalities. They also leverage the global event feature by only considering segments containing audio-visual events, i.e., they filter out background segments to compute the global feature. However, determining background segments often requires more supervision. In our work, we propose to use less supervision to fulfill the task. Different from [36], we introduce a recursive layer that can be stacked and therefore recursively fuses two uni-modal representations multiple times to obtain more robust representations.

Sound Localization is to associate certain regions in a video that has the corresponding sound with visual-aid. To this end, Hershey *et al.* [12] use a Gaussian process model to measure the mutual information of the audio and visual motion. Owens and Efros [23] propose a multi-sensory model that learns audio-visual correspondence in a self-supervised style to align the audio and visual frames and then localizes the sound source afterward. To investigate the correspondences between audio and visual components, Hu *et al.* [14] propose a deep multimodal clustering network that adds similar parts among two modalities to the final output. Zhao *et al.* [46] propose PixelPlayer that learns to locate image regions by leveraging large amounts of unlabeled videos. Arandjelovic and Zisserman [1] design two sub-networks that individually learn from audio tracks and image frames.

After learning, they fuse two branches to predict correspondence. Senocak *et al.* [29] develop a localization module that is based on the attention mechanism to capture the correlation between audio and visual features. The attention mechanism they adopt is quite plain where they transpose visual embedding and then multiply with audio embedding.

Multimodal Attention involves interaction at least two features from different modalities. In Video Question Answering (VQA), Lu *et al.* [20] propose a hierarchical attention technique that co-attends to the features extracted from text language modality and visual modality. Another work in VQA, Nguyen and Okatani [22] introduce a memory-based co-attention technique that enables dense interactions between the two modalities, and then both modalities contribute to the selection of the right answer. In emotion recognition, Zadeh *et al.* [44] exploit a small neural network that takes the concatenated cell states of three different LSTMs for language, audio, and visual components as input and then output the attended features. Wang *et al.* [37] present an attention gating mechanism where they try to learn a nonverbal shift vector by weighting features from different modalities. Different from the two aforementioned work that only perform attention operation once, we develop a joint co-attention mechanism that can be recursively performed.

Multimodal Fusion also known as the integration of information from multiple modalities [2], allows for more robust representations by leveraging multiple modalities and it can be categorized into two types: model-agnostic and model-based. Here, we only review the model-agnostic approaches, i.e., early fusion [31], late fusion [26, 31] and hybrid fusion [17, 40] as it is more related to our work. Early fusion combines low-level features of each modality while late fusion uses uni-modal decision values based on a fusion algorithm, e.g., averaging, weighting. Hybrid fusion, or double fusion, attempts to take advantage of both early and late fusion mechanisms. It has been widely used in the research of multimodal learning, e.g., multimodal speech recognition [39, 32] and multimedia event detection [17, 40, 15]. Besides a bigger picture of fusion strategies, there are many specific fusion strategies in terms of feature-level. Rahman *et al.* [25] adopt element-wise addition or multiplication, channel-wise concatenation, and fully-connected neural network to fuse information from three different modalities: language, audio, and visual modality.

3. Approach

In this section, we introduce the overall architecture of our proposed joint co-attention network for the supervised audio-visual event localization task layer by layer, as shown in Fig. 2. To start with the description, we first set forth the notations in Sec. 3.1, then the sequence feature re-representation layer is described in Sec. 3.2. Next, we introduce the proposed joint co-attention layer in Sec. 3.3. Lastly, we explain the final prediction layer in Sec. 3.4.

Table 1. Main symbols used throughout the paper.

Symbol	Definition
$\mathcal{S}_a/\mathcal{S}_v$	audio/visual sequence
s_a^t/s_v^t	audio/visual t-th segment-level feature
f_a^t/f_v^t	audio/visual feature after re-representation layer
$\mathbf{A}/\mathbf{V}/\mathbf{J}$	audio/visual/joint sequence-level feature
$\mathbf{C}_a/\mathbf{C}_v$	joint-audio/joint-visual affinity matrix
$d_a/d_v/d$	dimmesion of audio/visual/joint feature
ℓ	recursive times of joint co-attention layer
$\mathbf{H}_a/\mathbf{H}_v$	audio/visual feature after joint co-attention layer
$\mathbf{W}_{ja}/\mathbf{W}_{jv}$	parameters between ($\mathbf{J}$ and $\mathbf{A}$)/($\mathbf{J}$ and $\mathbf{V}$)
$\mathbf{W}_a/\mathbf{W}_v$	parameters for feature $\mathbf{A}/\mathbf{V}$
$\mathbf{W}_{ca}/\mathbf{W}_{cv}$	parameters for feature $\mathbf{C}_a/\mathbf{C}_v$
$\mathbf{W}_{h_a}/\mathbf{W}_{h_v}$	parameters for feature $\mathbf{H}_a/\mathbf{H}_v$

3.1. Notations

The symbols used throughout the paper are listed in Table 1. An Audio-Visual Event (AVE) is defined as an event that is both visible and audible [36]. As in [36, 38], for a given audio-visual video sequence $\mathcal{S} = (\mathcal{S}_a, \mathcal{S}_v)$, while $\mathcal{S}_a$ denotes the audio portion and $\mathcal{S}_v$ denotes the visual portion. The video sequence $\mathcal{S}$ is split into N non-overlapping yet continuous segments where each segment is typically one second long. For each segment, a label $y \in \{0, 1\}$ is given, while 0 indicates the segment is background and 1 indicates that is an AVE. The sequence features, i.e., $\mathcal{S}_a$ and $\mathcal{S}_v$ are extracted using a pre-trained CNN. We denote the extracted segment-level feature as s_a^t and s_v^t corresponding to the audio and visual modality respectively, where $t \in \{1, 2, \cdots, N\}$. Our network is built on the basis of fixed s_a^t and s_v^t.

3.2. Re-Representation Layer

Sequence representation contains temporal cues among the sequential stream, and LSTM has shown its superiority in learning those temporal cues. Therefore, we use the LSTM to modulate the sequence representations. Different from existing methods [19, 36], we add a residual embedding to the output of the LSTM in order to produce better representation. The structure of the proposed re-representation layer is shown in Fig. 2.

Audio Representation. In general, a sequence of audio feature $\mathbf{A}$ contains N continuous segments, i.e., $\{s_a^1, s_a^2, \cdots, s_a^N\}$, where each s_a^t is a 128×1 dimensional vector. We adopt Bi-directional LSTM [28] with residual embedding, in our case, concatenation, to learn the audio representation:

$$\overrightarrow{f_a^t} = \text{Bi-LSTM}(\begin{bmatrix} \overrightarrow{f_a^{t-1}} \\ s_a^t \end{bmatrix}), \tag{1}$$

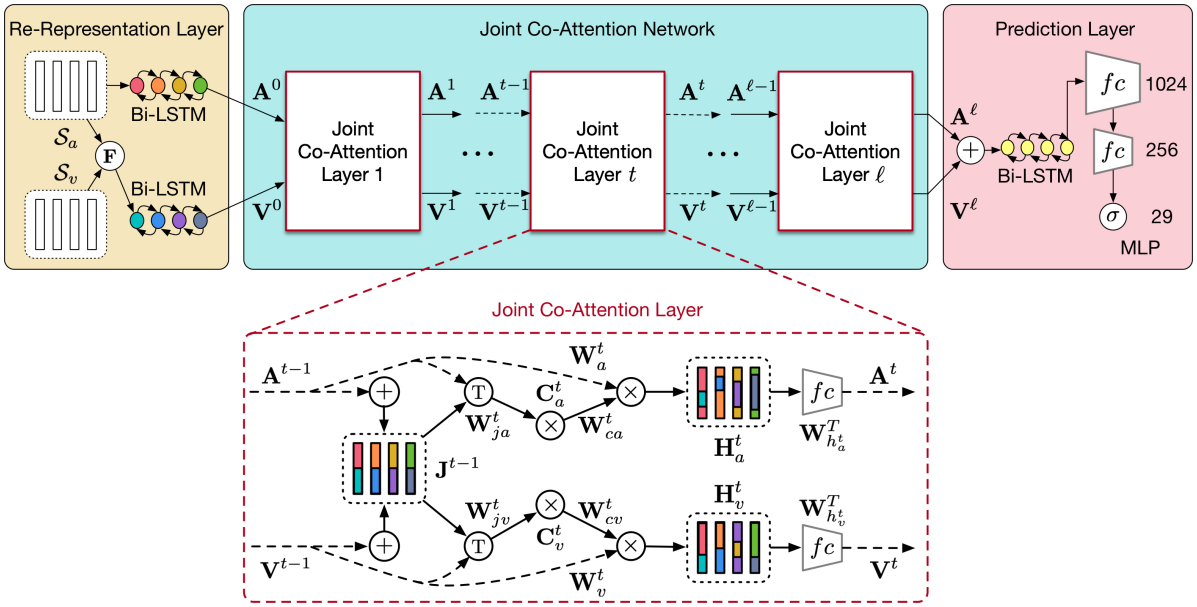

Figure 2. The overall structure of the proposed framework. We split it into three parts, i.e., sequence feature re-representation layer, joint co-attention network and category prediction layer. For the symbols, ⊕ denotes concatenation, Ⓕ denotes early fusion of audio feature and visual feature, Ⓢ denotes the softmax function, Ⓣ is transpose operator, and ⊗ is matrix multiplication operator.

$$\overleftarrow{f_a^t} = \text{Bi-LSTM}(\begin{bmatrix} \overleftarrow{f_a^{t+1}} \\ s_a^t \end{bmatrix}), \tag{2}$$

where the arrow indicates the direction of information flowing. Therefore, $f_a^t = \text{concat}(\overleftarrow{f_a^t}, \overrightarrow{f_a^t})$. We concatenate N segments along the time axis and create a new feature matrix, i.e., $\mathbf{A} = \text{concat}(f_a^1, \ldots, f_a^N) \in \mathbb{R}^{N \times d_a}$.

Visual Representation. Unlike audio features which are 1D features, visual features are 2D features extracted from image frames. This brings problems as the model needs to process two types of features with different dimensions simultaneously. Typically, the size of each visual feature is $512 \times 7 \times 7$ as in [36, 38]. If we simply conduct pooling in the height and width dimensions and reduce them into size 1 ($7 \rightarrow 1$), the performance of this reduction procedure can barely be guaranteed as the stride is big and may leave out useful information. Further study about the influence of applying different pooling methods is conducted in the ablation studies in Sec. 4.3. To smoothly reduce the dimension of raw visual features, we obtain the scaled dot-product of audio feature and visual feature for each segment. We denote the scaled dot-product as f_v^t. Then we follow a similar routine to encode the sequence of visual features like the audio sequence using LSTM with residual embedding. Consequently, we use a matrix for visual representation $\mathbf{V} = \text{concat}(f_v^1, \ldots, f_v^N) \in \mathbb{R}^{N \times d_v}$.

3.3. Joint Co-Attention Network

We now introduce the Joint Co-Attention (JCA) layer as shown in Fig. 2. The proposed joint co-attention layer attends to visual features and audio features simultaneously. It takes the audio representation $\mathbf{A}$ and the visual representation $\mathbf{V}$ as inputs and concatenates two representations as the joint representation $\mathbf{J}$. We employ $\mathbf{J}$ to co-attend to $\mathbf{A}$ and $\mathbf{V}$, respectively. It is worth noting that we only preserve $\mathbf{J} \rightarrow \mathbf{A}$ (i.e., joint feature attend to audio feature) and $\mathbf{J} \rightarrow \mathbf{V}$ (i.e., joint feature attend to visual feature), the inverse directions of $\mathbf{A} \rightarrow \mathbf{J}$ and $\mathbf{V} \rightarrow \mathbf{J}$ are abandoned for simplicity, which is different from the original co-attention mechanism [20]. One property of JCA is mutual attention, that is, it can attend to features from two different modalities simultaneously. Another special property of JCA is stackability, i.e., we can stack several JCAs so that we can recursively perform the process multiple times. Extensive experiments on different recursive times of the JCA unit are shown in Sec. 4.3.

Primary Idea for Joint Co-Attention. Recent studies [20, 22] explore the co-attention theory in Visual Question Answering (VQA). The text sequence representations and the visual sequence representations attend mutually to obtain new representations. Inspired by this, we explore a mode that allows representation from one modality not only attending to the other representation from the other modality but also attending to the representation from its original modality. Given audio representation $\mathbf{A} \in \mathbb{R}^{N \times d_a}$, and visual representation $\mathbf{V} \in \mathbb{R}^{N \times d_v}$, the joint representation $\mathbf{J} \in \mathbb{R}^{N \times d}$ is acquired by the concatenation of $\mathbf{A}$ and $\mathbf{V}$, i.e, $\mathbf{J} = \begin{bmatrix} \mathbf{A} \\ \mathbf{V} \end{bmatrix}$, where $d = d_a + d_v$. We take audio feature $\mathbf{A}^\ell$ as an example to elaborate the process of

joint co-attention. Here, we denote $\mathbf{A}^1$ as the initial state of audio feature and $\mathbf{A}^\ell$ as the audio feature after ℓ-th joint co-attention layer. First, the $(\ell-1)$-th layer's audio representation $\mathbf{A}^{\ell-1}$ is concatenated with $\mathbf{V}^{\ell-1}$ to obtain joint representation $\mathbf{J}^{\ell-1}$; next, we employ the $\mathbf{J}^{\ell-1}$ to attend to $\mathbf{A}^{\ell-1}$ and finally obtain the ℓ-th layer's audio feature $\mathbf{A}^\ell$. Following the similar rules, the new visual feature $\mathbf{V}^\ell$ is obtained.

Learning to Joint Co-Attend. Fusion is one of the key challenges for multimodal learning [2]. Following recent studies [20, 22] in VQA, we specifically derive the fusion to fit our audio-visual event localization task. After calculating the joint representation matrix $\mathbf{J}$, we use it to attend to different uni-modal representations via the following equation:

$$\mathbf{C}_a = \text{Tanh}\left(\frac{\mathbf{A}^{\mathrm{T}}\mathbf{W}_{ja}\mathbf{J}}{\sqrt{d}}\right), \tag{3}$$

where $\mathbf{C}_a$ is the joint-audio affinity matrix, T denotes transpose operation, and $\mathbf{W}_{ja} \in \mathbb{R}^{N\times N}$ is a learnable weight matrix ($\mathbf{W}_{ja}$ is implemented as fully-connected layer). Following the same rule, the joint-visual affinity matrix $\mathbf{C}_v$ can be written as

$$\mathbf{C}_v = \text{Tanh}\left(\frac{\mathbf{V}^{\mathrm{T}}\mathbf{W}_{jv}\mathbf{J}}{\sqrt{d}}\right), \tag{4}$$

where $\mathbf{W}_{jv} \in \mathbb{R}^{N\times N}$ is also a learnable weight matrix. After calculating the joint uni-modal affinity matrices $\mathbf{C}_a$ and $\mathbf{C}_v$, we then calculate the attention probabilities map $\mathbf{H}_a, \mathbf{H}_v$ of two modalities as, $\mathbf{H}_a = \text{ReLU}\left(\mathbf{W}_a\mathbf{A} + \mathbf{W}_{ca}\mathbf{C}_a^{\mathrm{T}}\right)$ and $\mathbf{H}_v = \text{ReLU}\left(\mathbf{W}_v\mathbf{V} + \mathbf{W}_{cv}\mathbf{C}_v^{\mathrm{T}}\right)$, where $\mathbf{H}_a \in \mathbb{R}^{k\times d_a}, \mathbf{H}_v \in \mathbb{R}^{k\times d_v}$ represent the attention probabilities map of audio modality and visual modality, respectively. $\mathbf{W}_a, \mathbf{W}_v \in \mathbb{R}^{k\times N}$, $\mathbf{W}_{ca}, \mathbf{W}_{cv} \in \mathbb{R}^{k\times d}$ are learnable weight matrices.

After obtaining the attention map $\mathbf{H}_a$ and $\mathbf{H}_v$, we recompute the new audio representation and new visual representation by

$$\mathbf{A}^\ell = g(\mathbf{A}^{\ell-1}, \mathbf{W}_{h_a^\ell}^{\mathrm{T}}\mathbf{H}_a^\ell), \tag{5}$$

$$\mathbf{V}^\ell = g(\mathbf{V}^{\ell-1}, \mathbf{W}_{h_v^\ell}^{\mathrm{T}}\mathbf{H}_v^\ell), \tag{6}$$

where $\mathbf{W}_{h_a^\ell}, \mathbf{W}_{h_v^\ell} \in \mathbb{R}^{k\times N}$ are learnable weight matrices in the ℓ-th layer. $\ell-1$ represents the features produced by the $\ell-1$-th layer. In our case, g is a summation function.

Fusion by Fusion. Multimodal fusion can generate more robust representation using the features from multiple modalities that are collected for the same phenomenon. Earlier studies [19, 36, 38] particularly exploit the method in an audio-visual dual-modality setting either directly fusing the features or using cross dot product operation. Different from them, we consider multimodal fusion as a recursive process, where we fuse audio representation $\mathbf{A}$ and visual representation $\mathbf{V}$ recursively to obtain more robust representations. Following Eq. (5) and Eq. (6), we generalize this recursive process as

$$\mathbf{A}^\ell = g(\cdots g(\mathbf{A}^0, \mathbf{W}_{h_a^1}^{\mathrm{T}}\mathbf{H}_a^1)\cdots, \mathbf{W}_{h_a^\ell}^{\mathrm{T}}\mathbf{H}_a^\ell), \tag{7}$$

$$\mathbf{V}^\ell = g(\cdots g(\mathbf{V}^0, \mathbf{W}_{h_v^1}^{\mathrm{T}}\mathbf{H}_v^1)\cdots, \mathbf{W}_{h_v^\ell}^{\mathrm{T}}\mathbf{H}_v^\ell), \tag{8}$$

where ℓ represents the amount of times that the joint co-attention is repeated. After fusing ℓ times, we will obtain two more robust representations for audio and visual modality, respectively.

3.4. Prediction Layer

The audio-visual event localization task aims to identify an AVE in a given video sequence and predict which category the AVE belongs to. Note that the input sequences of different categories and the backgrounds are heterogeneous. As a consequence, it is even harder to complete the task. Different from [38], we use less supervision by only considering event category labels. Before prediction, early fusion of two separate modalities is performed, and then two uni-modal representations are re-represented as $\mathbf{A}$ and $\mathbf{V}$. Next, following the fusion method in Sec. 3.3 of joint co-attention, we fuse two uni-modal representations multiple times to get $\mathbf{A}^\ell$ and $\mathbf{V}^\ell$. Finally, $\mathbf{A}^\ell$ and $\mathbf{V}^\ell$ are taken as input into the final category prediction layer:

$$\text{prediction} = \text{MLP}\Big(\text{Bi-LSTM}\Big(\begin{bmatrix}\mathbf{A}^\ell \\ \mathbf{V}^\ell\end{bmatrix}\Big)\Big). \tag{9}$$

where MLP denotes Multilayer Perceptron and Bi-LSTM is to modulate audio and visual representations jointly. In experiments, the MLP is implemented by using a two-layer fully-connected network embedded with 1,024/256 hidden units and a Sigmoid layer σ, as shown in Fig. 2. After that, the predicted category is the one that corresponds to the max value in the prediction vector. During training, we use the Multi Label Soft Margin loss function to optimize the entire network.

4. Experiments

4.1. Experimental Setup

Audio-Visual Event Dataset. The Audio-Visual Event (AVE) dataset by [36] is a subset of AudioSet [11]. It consists of $4,143$ video clips that involve 28 event categories. We adopt the split technology of [36] where train/validation/test sets are $3,309/402/402$ video clips, respectively. While training, the model has no access to the test portion to better evaluate the model's generalization ability. For the AVE dataset, it contains comprehensive audio-visual event types, in general, instrument performances, human daily activities, vehicle activities, and animal actions. To be more specific, for more detailed event

categories, take instrument performances as an example, AVE dataset contains accordion playing, guitar playing, and ukulele playing, etc. A typical video clip is 10 seconds long and is labeled with the start point and endpoint at the segment level to clarify whether the segment is an audio-visual event. Sample images and their attended images are shown in Fig. 3.

Evaluation Metrics. We follow [19, 36, 38] and adopt the global classification accuracy obtained from the last prediction layer as the evaluation metric. For an input video sequence, our goal is to predict the category label for each segment. It is worth noting that the background category contains 28 backgrounds since each event category can have its own background so that it is hard to predict.

Experimental Details. Following [36, 38], we adopt pre-trained CNN models to extract features for each audio and visual segment. Specifically, we exploit the VGG19 [30] network pre-trained on ImageNet [5] as the backbone to extract segment-level visual features. Meanwhile, for the audio segment, we extract the segment feature using a Vggish network [13] which is pre-trained on AudioSet [11]. For a fair comparison, we use the same extracted features (i.e, audio and visual features) as used in [36, 38]. In the training stage, the only supervision we exploit is the annotation labels for the temporal segments.

4.2. Comparison with Existing Methods

State-of-the-Art Comparison. Results compared with the leading methods are reported in Table 2. We take a similar model architecture as in [36] and run single modality models as our baselines, which only take audio features or visual features during the experiments. First, to validate the proposed method can enable efficient interactions between audio features and visual features, we compare with a state-of-the-art temporal labeling network, i.e, ED-TCN [18], which can integrate information from multiple temporal segments. Next, to verify the effectiveness of our fusion strategy of audio feature and visual feature, we compare with two methods, i.e, Audio-Visual [36] and AVSDN [19]. Both methods utilize a straightforward fusion strategy, where fuses the audio and visual features out of LSTMs by concatenation. Lastly, to evaluate that our method is tolerant with less supervision, we compare our method with DAM [38], which needs additional supervision to exclude event-irrelevant segments during training.

Comparison Analysis. Due to the absence of interactions between audio modality and visual modality, our proposed model can easily surpass the performance of the baselines. In addition, by comparing with ED-TCN, our model enables more effective interactions between two modalities. Thus, it can be testified that interactions or fusion can boost the task performance and our model is more superior on enabling interactions between two different modalities. Unsurprisingly, by fusing the two different features using our joint co-attention mechanism, our model outperforms Audio-Visual and AVSDN using a plain fusion strategy. Moreover, even without additional effort to exclude event-irrelevant segments, our model can learn useful representations from noisy inputs and contribute to better performance.

Table 2. Results of comparisons with the state-of-the-art methods on the AVE dataset. For a fair comparison, * is obtained by exploiting the same pre-trained audio and visual features. While the task is hard, it can still be observed that our model outperforms the existing methods.

Method	Accuracy (%)
Audio_Only (Vggish [13])	59.5
Visual_Only (Vgg19 [30])	55.3
ED-TCN [18]	46.9
Audio-Visual [36]	71.4
AVSDN* [19]	72.6
Full-Audio-Visual [36]	72.7
DAM [38]	74.5
Ours	**76.2**

Table 3. Ablation studies on the proposed framework. Uni-modal Bi-LSTM is the LSTM in sequence feature re-representation layer while Joint Bi-LSTM is the one in the prediction layer. * denotes we remove the residual embedding of LSTMs while † denotes that we adopt the primary co-attention mechanism into the proposed framework.

Model	Accuracy (%)
Ours w/o Uni-modal Bi-LSTM	74.5
Ours w/o Joint Bi-LSTM	74.9
Ours w/o Residual Embedding*	75.2
Ours w/ GRU [4]	75.3
Ours w/ Average Pooling	75.1
Ours w/ Max Pooling	75.0
Ours w/ Co-Attention† [22]	75.4
Ours w/ Joint Co-Attention	**76.2**

Table 4. Results of different fusion strategies to generate joint representation $\mathbf{J}$.

Strategy	Accuracy (%)	Params $\times 10^6$
Addition	75.0	**22.67**
Multiplication	74.6	**22.67**
Concatenation	75.2	22.72
Addition + FC	75.5	22.78
Multiplication + FC	75.3	22.78
Concatenation + FC	**76.2**	22.83

4.3. Ablation Study

Framework Decoupling. We break down the proposed framework and evaluate them separately in different settings, as shown in Table 3. For the Bi-LSTM, we define it as two types, one in sequence feature re-representation as uni-modal Bi-LSTM while the other in the prediction layer as joint Bi-LSTM. Experiments on two Bi-LSTMs are denoted as 'Ours w/o Uni-modal Bi-LSTM' and 'Ours w/o Joint Bi-LSTM, respectively.

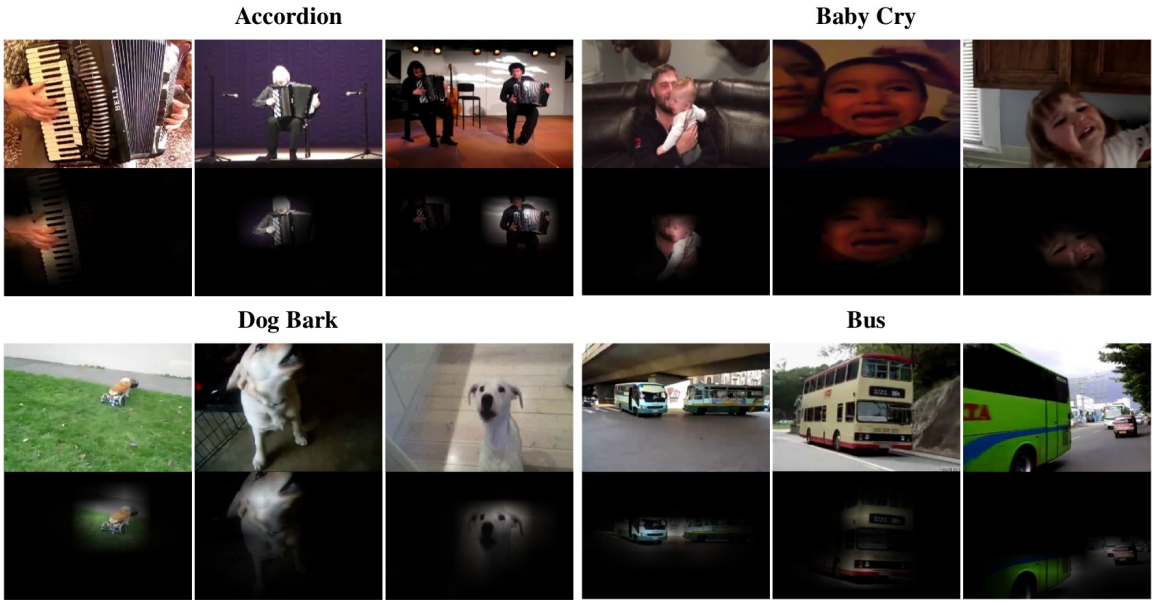

Figure 3. Visualizing attention mask of the proposed joint co-attention mechanism on four categories of the AVE dataset.

Table 5. Variations on the proposed JCA architecture. Unlisted value are identical to those of the first row of the model. Besides accuracy, we also calculate the parameters of each setting.

	d_a	d_v	ℓ	Accuracy (%)	Params $\times 10^6$
(A) 256×1024	256	1024	4	75.4	50.7
(B) 512 × 512	512	512	1	75.1	14.9
			2	75.4	17.5
			3	75.6	20.1
			4	**76.2**	22.8
			5	75.6	25.4
(C) 256 × 256	256	256	2	74.8	4.6
			3	75.0	5.2
			4	75.1	5.9
			5	74.8	6.6
(D) 128 × 128	128	128	4	73.8	**1.6**

Moreover, GRU [4] is used as an alternative to Bi-LSTM for further investigation, denoted as 'Ours w/ GRU'. For early fusion, we evaluate two direct pooling methods i.e., global average pooling and global max pooling, denoted as 'Ours w/ Average Pooling' and 'Ours w/ Max Pooling', respectively. Lastly, the 'Ours w/ Co-Attention' represents that we replace joint co-attention with the original co-attention [20]. Our full model is denoted as 'Ours w/ Joint Co-Attention'.

Framework Analysis. Results are showed in Table 3. First, the overall performance of the proposed framework outperforms the state-of-the-art method [38] which needs additional supervision. Among all the observed declines, Bi-LSTM has the highest impact. That confirms the effectiveness of the Bi-LSTM part. For alternatives to early fusion, neither the global average pooling nor the global max pooling surpasses our full model.

Among the experiment results with two different co-attention mechanisms, i.e., original co-attention method [22] and our joint co-attention method, our joint co-attention method excels the original co-attention method which follows a dual-modality mutual attending way (visual features attend to audio features and audio features attend to visual features). By not only attending to the corresponding modality but also the modality of itself, our proposed joint co-attention method performs better in the audio-visual fusion task. To sum up, the ablation studies demonstrate the efficiency of our proposed framework.

Studies on Different Fusion Strategies. To further investigate how different fusion strategies used to produce joint representation **J** can influence the performance of the proposed model, we exploit various fusion strategies as variations of our proposed model. (i) Element-wise addition; (ii) Element-wise multiplication; (iii) Channel-wise concatenation; (iv) Fully-connected neural network (FC).

Results are showed in Table 4. It can be witnessed that directly making concatenation, addition or multiplication impair the performance of the fusion representation while introducing FC can slightly increase numbers of the parameters (less than 0.7%), but the performance can increase by 2.1%.

Studies on Recursive Times ℓ, and Dimensions of d_a and d_v. To further investigate the proposed framework, we vary the proposed model in different ways and then evaluate the accuracy under each circumstance. The results are presented in Table 5.

The ℓ denotes the times of JCA that are recursively performed whereas d_a and d_v denote the dimension of audio feature and visual feature, respectively. We observe

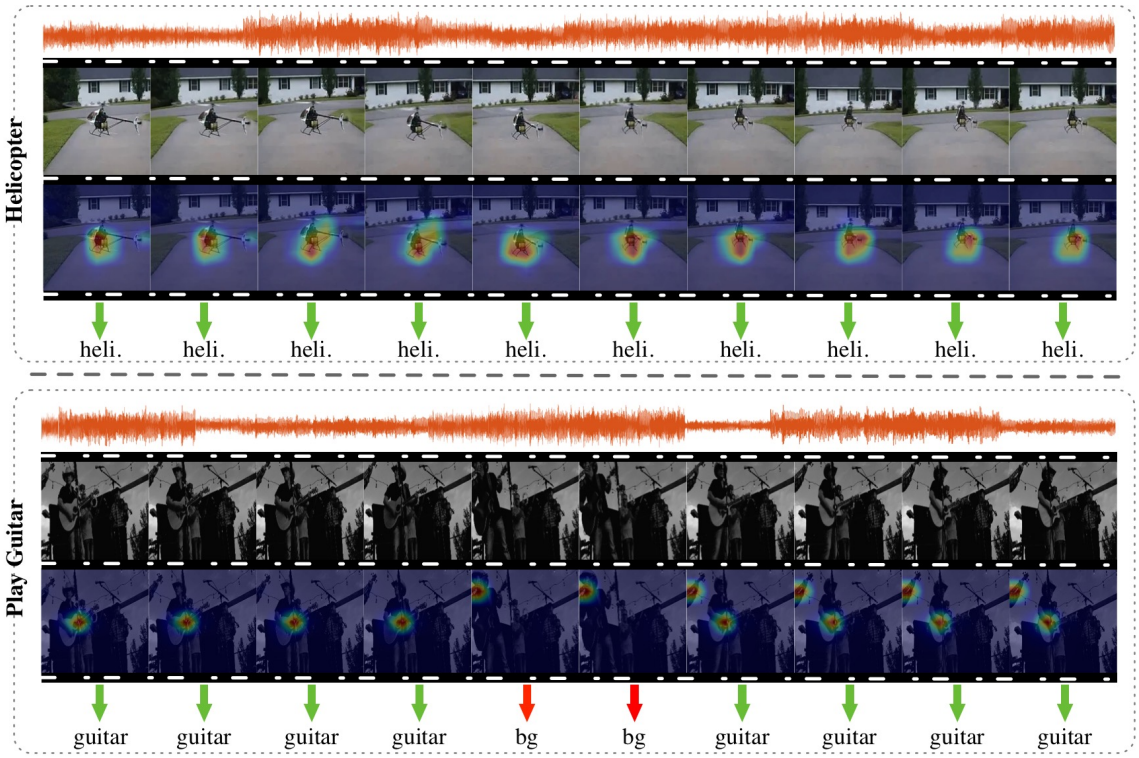

Figure 4. Two qualitative results on audio-visual localization task. The first example is helicopter hovering, i.e, 'heli.' is the abbreviation for helicopter for better layout; while second example is playing guitar, i.e., 'guitar' for short, 'bg' denotes 'background'. The green arrow represents the correct prediction whereas the red arrow denotes the wrong prediction. To visualize where they attend to, we generate images with their corresponding attention map. *Best viewed in color*.

that reducing the dimensions of the input features (d_a and d_v) hurts the model's performance, which suggests high-dimensional features may be suitable for the fusion. However, features with extremely high dimension would bring a lot of computation. In practice, one should make a trade-off here. If we only look at row (B) or row (C), it is easy to find that as the recursive times of JCA increase, the performance improves. This also validates our motivation that repeating the fusion process helps our model to learn more robust representations.

4.4. Qualitative Evaluation

In this section, we show some qualitative results of our proposed framework in Fig. 4. For each row in Fig. 4, the left is the category of this audio-visual event; the top content is the waveform of input audio sequence; the middles are raw frames and frames with attention map of the input video sequence; the bottom is the audio-visual event prediction.

Among the two instances in Fig. 4, the second instance is much harder as the scene is more complicated where different people are playing different instruments. In the beginning, the proposed network predicts well. However, as the singer changing his posture, the guitar can hardly be seen even with our eyes. Therefore, the network fails to predict it as playing guitar. Surprisingly, as the singer turns back to the front, our network works again, and it marks two guitars in the picture even the other guitar is indistinct. More results are shown in Fig. 3. We can see that the proposed co-attention model adaptively captures different sound sources in different semantic regions, such as accordion, crying boy/girl/babies, barking dog, horning bus, ukulele, etc.

5. Conclusion

In this paper, we investigate an interesting problem on deep audio-visual learning for the AVE task. To better cope with this multimodal learning task, we propose a novel joint co-attention mechanism with double fusion. To the best of our knowledge, this is the first time of applying the co-attention mechanism into the audio-visual event localization task. The integration with double fusion leading to better representations for the AVE task by co-attending to both audio and visual modalities. Moreover, experimental results on the AVE dataset have confirmed the superiority of the proposed framework.

Acknowledgement: This research was partially supported by NSF CSR-1908658 and NeTS-1909185. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

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

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 2018.

[3] Xinya Chen, Yanrui Bin, Changxin Gao, Nong Sang, and Hao Tang. Relevant region prediction for crowd counting. *Elsevier Neurocomputing*, 2020.

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

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Lei Ding, Hao Tang, and Lorenzo Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE TGRS*, 2020.

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

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

[10] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019.

[11] 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.

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

[13] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[14] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019.

[15] Lu Jiang, Alexander G Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, 2012.

[16] Faheem Khan and Ben Milner. Speaker separation using visually-derived binary masks. In *AVSP*, 2013.

[17] Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 2014.

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

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

[20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.

[21] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.

[22] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*, 2018.

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

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

[25] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, 2019.

[26] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *ACII*, 2011.

[27] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised segmentation and source separation on videos. In *CVPR Workshops*, 2019.

[28] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE TSP*, 1997.

[29] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.

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

[31] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005.

[32] Felix Sun, David Harwath, and James Glass. Look, listen, and decode: Multimodal speech recognition with images. In *SLT Workshop*, 2016.

[33] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020.

[34] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019.

[35] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*, 2019.

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

[37] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, 2019.

[38] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.

[39] Zhiyong Wu, Lianhong Cai, and Helen Meng. Multi-level fusion of audio and visual features for speaker identification. In *ICB*, 2006.

[40] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, 2015.

[41] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 2018.

[42] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019.

[43] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017.

[44] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018.

[45] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

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