This WACV 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

See the Sound, Hear the Pixels

Janani Ramaswamy, Sukhendu Das
Visualization and Perception Lab,
Dept. of Computer Science and Engineering, IIT Madras, India
janani@cse.iitm.ac.in, sdas@iitm.ac.in

Abstract

For every event occurring in the real world, most often a sound is associated with the corresponding visual scene. Humans possess an inherent ability to automatically map the audio content with visual scenes leading to an effortless and enhanced understanding of the underlying event. This triggers an interesting question: Can this natural correspondence between video and audio, which has been diminutively explored so far, be learned by a machine and modeled jointly to localize the sound source in a visual scene? In this paper, we propose a novel algorithm that addresses the problem of localizing sound source in unconstrained videos, which uses efficient fusion and attention mechanisms. Two novel blocks namely, Audio Visual Fusion Block (AVFB) and Segment-Wise Attention Block (SWAB) have been developed for this purpose. Quantitative and qualitative evaluations show that it is feasible to use the same algorithm with minor modifications to serve the purpose of sound localization using three different types of learning: supervised, weakly supervised and unsupervised. A novel Audio Visual Triplet Gram Matrix Loss (AVTGML) has been proposed as a loss function to learn the localization in an unsupervised way. Our empirical evaluations demonstrate a significant increase in performance over the existing state-of-the-art methods, serving as a testimony to the superiority of our proposed approach.

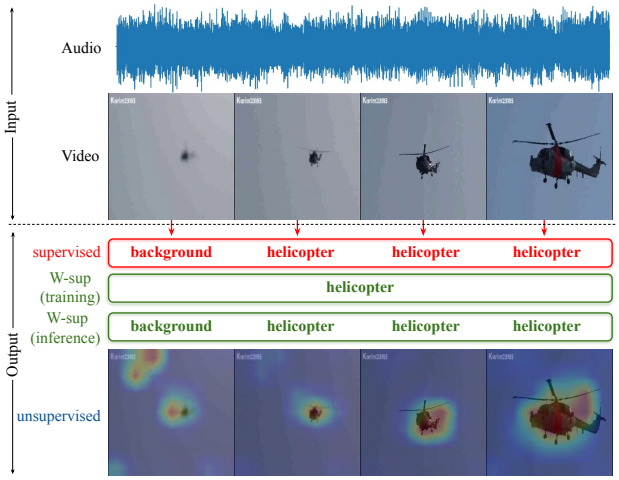

Figure 1. **Example of sound source localization.** Three tasks depicted for a video split into shorter non-overlapping segments: supervised event localization (shown in red) which predicts event labels for each video segment, weakly supervised event localization (shown in green) which is trained by giving only one event label for the whole video but is made to predict event labels for every segment in the video during inference, and unsupervised sound source localization which gives attention maps without using any event labels. The attention maps are inaccurate in the absence of a joint audio-visual event (*i.e.* not properly audible or visible or both) in a segment, as evident from the attention map of the left-most frame of the bottom row (unsupervised case).

1. Introduction

Visual events are most often correlated with sound leading to better audio and visual comprehension of the scene. They can be considered as two different outlooks [2] of the same data. Hence, an integrated analysis of the two modalities can provide rich spatial and temporal cues to localize sound source in visual scenes. Humans are empowered to learn this implicit mapping [2, 39], which is performed unconsciously due to the ubiquitous availability of audio-visual examples around them. They tend to learn strong correlations between the two sensory streams from a very early stage of life [28, 37]. Can this technique be leveraged to jointly model an audio-visual learning for event localization? Can this correlation be captured to make machines learn to understand real-life events using high-level semantic representations of the two modalities? Given the availability of a large number of unlabeled videos, is it possible to obtain supervision from only the input videos without any event labels? These questions if answered, open avenues to a wide range of areas like audio-visual scene understanding, audio-based saliency detection, segmentation, surveillance, lip-reading for surveillance oriented applica-

tions and audio-based video captioning. However, this is a highly challenging task as audio is not as semantically close to images/videos as like text [3]. Hence, combining audio with vision is tougher than combining text with vision. In addition, audio not being perfectly in sync with video, presence of ambient sound like breeze, as well as the object producing the sound being momentarily occluded in the video, are other important factors that complicate the task of localization and scene understanding.

Although there have been several attempts on audio-visual fusion [6, 9, 16, 20, 21] prior to the deep learning era, the release of datasets like SoundNet [4] and Audioset [14] fueled more interest in the domain and an unprecedented spike in research has been witnessed ever since. Following that, there has been some work on sound source localization [2, 3, 11, 23, 24, 34, 39, 44, 47, 53], most of which work in a self-supervised fashion making use of two sub-networks to extract features from the audio and video streams separately and one module to enable fusion of these features. Output of fusion indicates whether the audio corresponds to that particular video or not. Attention maps are obtained as intermediate results of the process. However, the drawbacks that have been generally observed are: 1) some works [3, 53] have not showcased their observations across diverse unconstrained videos but restricted to a specific domain like music; 2) a few works [2, 3, 39] consider only one frame representing a full video thus not taking the temporal information into consideration; 3) works [24, 44] that deal with unconstrained videos have not shown sound localization in an unsupervised setup; 4) almost all the works use simple fusion strategies like element-wise multiplication, addition and feature concatenation to fuse audio and visual content which is not rich enough to find high-level associations between the two modalities.

This instilled our interest to perform sound localization in unconstrained videos using both spatial and temporal information. The goal of this paper is to jointly model audio and visual information from unconstrained videos to get compact feature representations that can localize the object producing the sound in a scene and thereby perform event localization, when provided with event labels for supervision. Three types of learning: supervised, weakly supervised and unsupervised, have thus been systematically investigated (as shown in figure 1) for this task using the same algorithm with minor changes, thereby exhibiting the scalability and robustness of the proposed method.

We propose a novel method which uses two blocks for processing: Audio Visual Fusion Block (AVFB) and Segment-Wise Attention Block (SWAB). AVFB uses a hybrid of neural network (LSTM) [25, 26, 32, 33, 49] and multi-modal bilinear pooling [7, 10, 22, 41, 42, 50] to fuse the feature representations obtained after extraction using pre-trained CNNs from the two modalities (that is, from audio of single-channel and video which are broken into segments of 1 second each). SWAB uses the audio-assisted visual features coming from the fusion block (AVFB) and the audio features, along with the global information from the respective modalities, to localize sound source in the scene by providing segment-wise attention. The feature representations from AVFB and SWAB are then aggregated together and fed into fully connected layers to obtain the event labels (in case of supervised and weakly supervised learning tasks). We also propose a novel loss function, AVTGML (Audio Visual Triplet Gram Matrix Loss) to localize sound source without any event labels. Our key contributions are summarized as follows: i) We propose an Audio Visual Fusion Block (AVFB) which effectively fuses the features extracted from audio and video streams to provide audio-assisted visual features, where audio helps in attending to specific regions in the video by providing corresponding weightage to the spatial regions; ii) We propose a Segment-Wise Attention Block (SWAB) which combines global information of the two modalities with audio-assisted visual features and audio features correspondingly such that it weighs the segments in the video according to the importance of segments in the audio; iii) We propose an Audio Visual Triplet Gram Matrix Loss (AVTGML) function to localize sound source in an unsupervised way. The fact that this loss takes the dynamic inter-segment relationship into account makes it different from other triplet loss functions used for this application [39]; iv) Our experimental results on the AVE dataset [44] demonstrate that our method significantly outperforms the existing state-of-the-art methods.

2. Related Work

Though there has been prolific research on audio-visual cross-modal analysis for different applications, we focus only on topics that are more relevant to our work of sound source localization. We also describe how our work differs from traditional methods, recent deep learning based sound source localization and other closely related audio-visual representation learning techniques.

Comparison with human sensory perception The motivation of making machines learn aligned representations of audio and video stems from the way humans integrate multi-sensory information. The work done in [28, 37] show how this fusion arises from a very early stage of life. An evident example of such audio-visual integration in humans is McGurk effect [27] which elucidates the importance of vision in speech perception. Likewise, Sekular *et al*. [38] reveals that sound can also alter visual perception. We take inspiration from this human mechanism to develop an algorithm that integrates audio and visual information to efficiently localize sound for better scene perception.

Traditional audio-visual correspondence learning Even before deep learning made its feat in the vision and

audition domain, few works attempted to combine the two modalities using conventional low-level features [6, 9, 16, 20, 21]. One such early work [16] relied on audio-visual synchrony for sound source localization which modeled the signals as a non-stationary Gaussian process. The joint distribution of audio & video signals was learnt in [9] using a non-parametric approach by projecting data into a low-dimensional subspace. Spatial sparsity constraints were used to achieve audio-visual correlation in [21]. Barzelay and Schechner [6] explored temporal coincidences (motion cues) between the 2 modalities to accomplish cross-modal association. Likewise, Izadinia *et al*. [20] used Canonical Correlation Analysis (CCA) to capture audio-visual correlation for detecting and segmenting moving objects. In contrast, our work doesn't use low-level handcrafted features (*e.g*. gradient of intensity values) and exploits deep learning instead, to automatically learn the fusion between audio and video, accommodating as much variability in data as possible without imposing any constraints, thereby ensuring generalizability and robustness.

Deep learning based audio-visual correspondence learning There has been a lot of focus in research towards audio visual cross-modal analysis where audio and visual information are considered supervisory to each other and hence are used for self-supervision. Senocak *et al*. [39] propose an unsupervised method for sound source localization and show that a bit of prior knowledge helps in improving the model performance. Arandjelovic and Zisserman [2, 3] also learn a good audio-visual correspondence using self-supervised learning. Along similar lines, audio source separation using both audio and visual signals have been attempted in [12, 13, 29, 48, 52, 53]. Tian *et al*. [44] and Lin *et al*. [24] perform audio-visual event localization using supervised and weakly supervised learning. Deep multimodal clustering is employed to get efficient audio-visual correspondence in [19]. Unlike these works, our method provides efficient fusion of audio and visual information from unconstrained videos by also providing segment-wise attention leading to superior performance.

Other audio visual based applications Aytar *et al*. [4] propose a student-teacher network to transfer discriminative visual information to sound modality. Owens *et al*. [30] use a recurrent neural network to synthesize sound from silent videos. Visual representations are learnt using ambient sound in [31]. Another prior work [8] attempts to perform lip reading using ConvNet architecture. There has also been a boom in research which involves combining the three modalities: text, visual content and audio. A few such recent works [1, 18, 36] make use of both audio and visual features to answer users' questions about dynamic scenes using natural language (scene-aware dialog). Harwath *et al*. [15] use unsupervised learning to analyze associations between image scenes and spoken audio captions. Tian *et al*. [43] develop an audio-visual based video captioning network in an attempt to combine text with audio and visual sequences. Vision, sound and text have also been used collectively to achieve an aligned representation for cross-modal retrieval and other similar tasks in [5]. Contrary to all these methods, our work focuses solely on using visual content and its audio counterpart to precisely localize sound source in a visual scene.

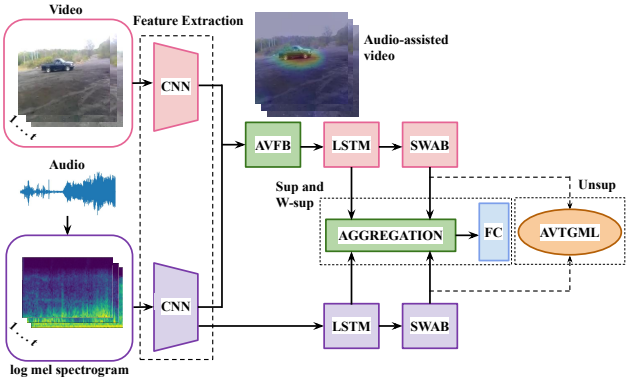

Figure 2. **The overall architecture** of our proposed model to perform sound source localization in a visual scene using three different types of learning: supervised, weakly supervised and unsupervised. AVFB and SWAB refer to the proposed 'Audio Visual Fusion Block' (for efficiently fusing features from the two modalities) and 'Segment-Wise Attention Block' (to provide segment-wise attention apart from region-wise attention) respectively. In case of unsupervised learning, the feature representations from SWABs are used in the proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function to get the attention maps.

3. Proposed Algorithm

The proposed algorithm (as shown in figure 2) for sound source localization consists of the following sequence of steps: 1) Extract features from vision and sound modalities using CNNs; 2) Fuse the extracted features from the two modalities using Audio Visual Fusion Block (AVFB) to get the audio-assisted visual features that contain the attention to be given to each spatial region in each segment of the video; 3) Feed the audio-assisted visual features and audio features respectively into LSTMs to model their temporal dependencies; 4) Give the outputs of the two LSTMs to their respective Segment-Wise Attention Block (SWAB) to ensure that attention is given not only to the spatial region in each segment, but also to the segments of both the modalities themselves; 5) Aggregate the feature representations of both the modalities; 6) Feed the result of aggregation to a series of fully connected layers to get segment-level labels (supervised) or video-level labels (weakly supervised); 7) In the case of unsupervised learning, apart from the visual and its audio counterpart, extract the features of a negative au-

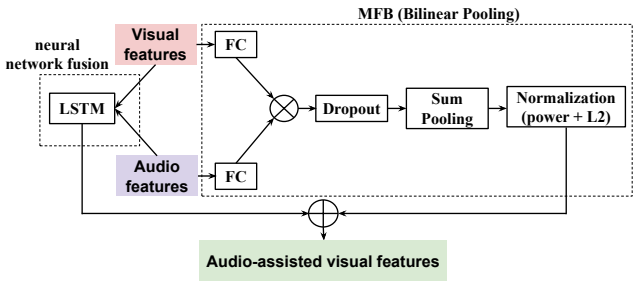

Figure 3. **Audio Visual Fusion Block (AVFB)**. This integrates neural network based fusion with a bilinear model (MFB) [50] to fuse the visual and audio features extracted from CNNs. $\otimes$ and $\oplus$ refer to element-wise multiplication and addition respectively.

dio sample (that is, an audio belonging to any other random video) as well, get the corresponding outputs from SWAB and feed them to the proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function. In this way, the same architecture is used with minor changes to accomplish all the three types of learning tasks.

3.1. Notations

Based on [44], we consider unconstrained videos which have audio-visual events that are both visible and audible. Let the video and audio channels be represented as $\{\mathcal{VC}, \mathcal{AC}\}$. Each video of ten seconds duration, is split into $\mathcal{T}(=10)$ non-overlapping segments of length one second each. We denote each such tuple of visual and audio segments as $\{\mathcal{V}_t, \mathcal{A}_t\}_{t=1}^{\mathcal{T}}$. Following [44], the event label for the video segment is denoted as $y_t = \{y_t^c | y_t^c \in \{0,1\}, c = 1, ..., C, \sum_{c=1}^{C} y_t^c = 1\}$. Here, C denotes the total number of event categories plus one label for background. Audio and visual features are extracted from these raw segments using pre-trained CNNs [17, 40], which are denoted as $\{F_t^v, F_t^a\}_{t=1}^{\mathcal{T}}$ where $F_t^v \in \mathbb{R}^{d_v \times S}$ and $F_t^a \in \mathbb{R}^{d_a}$. Here, d_v is the number of CNN feature maps, S is the vectorized spatial dimension of each feature map and d_a refers to the dimension of audio features. Similar to [44], the rest of our architecture is built on top of these local features.

3.2. Audio Visual Fusion Block (AVFB)

Once the features are extracted from pre-trained CNNs [17, 40], they are fed to the AVFB (shown in figure 3). One time-step is shown in case of LSTM based fusion while the MFB module is shared for all time steps in figure 3. This block combines two fusion strategies to effectively fuse the audio and visual features as follows:

LSTM based fusion: The audio and visual features are initially taken to a common embedding space so that they can be concatenated and then fed to an LSTM for fusion. To achieve this, we utilise global average pooling so that the visual feature obtained from a pre-trained CNN, $F_t^v \in \mathbb{R}^{d_v \times S}$, is converted to a feature representation $l_t^v \in \mathbb{R}^{d_e}$, and similarly the audio features $F_t^a \in \mathbb{R}^{d_a}$ are passed through dense layers to convert them to feature $l_t^a \in \mathbb{R}^{d_e}$, where d_e is the dimension of the common embedding space. For every segment, features l_t^v and l_t^a are concatenated sequentially as $([l_1^v + l_1^a, l_1^a], [l_2^v + l_2^a, l_2^a], ..., [l_{\mathcal{T}}^v + l_{\mathcal{T}}^a, l_{\mathcal{T}}^a])$, so that the visual content obtains the necessary time-stamped information from its audio counterpart for efficient fusion. Given an input $[l_t^v + l_t^a, l_t^a]$, the update of hidden and cell states produced by LSTM is represented as:

$$h_t^{'}, c_t^{'} = LSTM([l_t^v + l_t^a, l_t^a], h_{t-1}^{'}, c_{t-1}^{'}) \tag{1}$$

We shall interchangeably use $h_t^{'}$ and LF_t^v, which represents the output of LSTM-based fusion between audio & visual features. This fusion is along the lines of [25, 26, 32, 33, 49] who also employ LSTM based fusion, but for visual question answering.

Adapting Multi-modal Factorized Bilinear Pooling (MFB) for audio-visual fusion: A standard bilinear model can be represented as:

$$\mathbf{y}_i = \mathbf{x}_1^T W_i \mathbf{x}_2 \quad i = 1, ..., p \tag{2}$$

where, $\mathbf{x}_1 \in \mathbb{R}^{d_{\mathbf{x}_1}}$ and $\mathbf{x}_2 \in \mathbb{R}^{d_{\mathbf{x}_2}}$ are feature vectors from two different modalities to be fused, while $W = [W_1, ..., W_p] \in \mathbb{R}^{d_{\mathbf{x}_1} \times d_{\mathbf{x}_2} \times p}$ is the projection matrix that is learnt to get a p-dimensional output $\mathbf{y}$. Due to colossal amount of parameters and high computational cost in this model, Yu *et al*. [50] proposed Multi-modal Factorized Bilinear Pooling (MFB), where W is factorized into two low-rank matrices. Absorbing this idea from [50] and adapting it to our case of audio-visual fusion, we get:

$$\tilde{z}_t = SumPooling(U^T F_t^v \circ V^T F_t^a, q) \tag{3}$$

$$z_t^{'} = sign(\tilde{z}_t)|\tilde{z}_t|^{0.5}; z_t = z_t^{'}/||z_t^{'}|| \tag{4}$$

where, $U \in \mathbb{R}^{d_v \times (qp)}$ and $V \in \mathbb{R}^{d_a \times (qp)}$ are two low-rank matrices obtained from W which are learnt, q represents latent dimensionality and $\circ$ refers to Hadamard product. The output of MFB module (z_t) is used to estimate the attention weight vector λ_t providing attention over the spatial regions based on the weightage provided by its audio counterpart, as:

$$\lambda_t = Softmax(W_z z_t + b_z) \quad t = 1, ..., \mathcal{T} \tag{5}$$

where, W_z and b_z are learnable parameters. The audio aware visual vector for each time step t is computed as:

$$MF_t^v = MFB(F_t^v, F_t^a) = \sum_{i=1}^{S} \lambda_t^i F_t^{v^i} \tag{6}$$

This fusion based attention gives the relevance of each spatial grid (in each frame) to the audio, leading to a better understanding of the event. Finally, by combining the outputs

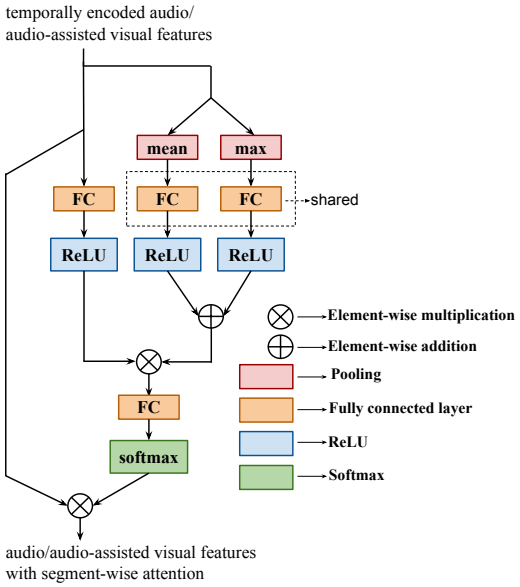

Figure 4. **Segment-Wise Attention Block (SWAB)**. This helps to learn the degree of attention required in each segment in addition to the spatial attention provided by AVFB. The module takes as input, feature encodings from LSTMs (see figure 2) and generates segment-wise attention as output for aggregation and fusion.

of the above two fusion methods, the audio-assisted visual features are obtained as:

$$IF_t^v = LF_t^v + MF_t^v \tag{7}$$

Hence, the advantages of both LSTM and bilinear model based fusion have been exploited by our proposed framework.

3.3. Modeling temporal dependencies

Since audio and video are both sequential in nature, their temporal dependencies are better encoded using two separate LSTMs, given an input X_t (as either IF_t^v or F_t^a), as:

$$h_t^m, c_t^m = LSTM(X_t, h_{t-1}^m, c_{t-1}^m) \tag{8}$$

where, m can be a (audio) or v (audio-assisted visual) depending on what X_t is.

3.4. Segment-Wise Attention Block (SWAB)

Apart from spatial attention provided by AVFB, it is also important to give weightage to each segment of audio and visual content, because not all segments provide equal amount of information about an event. Certain time steps reveal more information about an event than other time steps. The sound at a particular segment might give stronger clues about an event compared to other segments and likewise for its visual counterpart. Hence, giving equal attention to all segments would seem unreasonable. In order to provide segment-level attention (shown in figure 4), an overall knowledge of all segments is required beforehand.

Pooling is used to get this global information. As quoted in [45], although mean pooling succeeds in giving an overall intuition of what the content is about, it tends to leave out fine discriminative details. Hence a combination of mean and max pooling ensures in capturing the overall content without missing any major detail. The empirical evidence of the superior performance of this combination is discussed later. Let the outputs of mean and max pooling layers be denoted as R_{ave}^m and R_{max}^m respectively. The process of providing segment-wise attention can be mathematically formulated as follows:

$$\begin{aligned} g_{ave}^m &= relu(W_g^m R_{ave}^m + b_g^m) \\ g_{max}^m &= relu(W_g^m R_{max}^m + b_g^m) \\ \bar{g}_t^m &= relu(W_{\bar{g}}^m h_t^m + b_{\bar{g}}^m) \\ \tilde{g}^m &= g_{ave}^m + g_{max}^m; P_t^m = \bar{g}_t^m \otimes \tilde{g}^m \end{aligned} \tag{9}$$

where, $\otimes$ refers to element-wise multiplication. The segment-attention weights are then computed as:

$$\alpha_t^m = Softmax(W_\alpha^m P_t^m + b_\alpha^m) \tag{10}$$

Finally, scaling the features in each segment h_t^m by the corresponding attention weights α_t^m, provides the features with varied segment-wise importance in a video, as:

$$s_t^m = \alpha_t^m \otimes h_t^m \tag{11}$$

3.5. Aggregation

We have two feature representations in each modality: region-wise and segment-wise attended features. A similarity measure, motivated from [51], is calculated between these two types of feature representations as:

$$\begin{aligned} sim_t &= (h_t^a - h_t^v) \circ (s_t^a - s_t^v) \\ &= (h_t^a \circ s_t^a - h_t^v \circ s_t^a) + (h_t^v \circ s_t^v - h_t^a \circ s_t^v) \\ &= (c_{11}F_t^a \circ c_{12}F_t^a - c_{22}F_t^v \circ c_{12}F_t^a) \\ &+ (c_{22}F_t^v \circ c_{21}F_t^v - c_{11}F_t^a \circ c_{21}F_t^v) \end{aligned} \tag{12}$$

where, c_{11}, c_{12}, c_{21} and c_{22} are the coefficients that modify the initial features extracted from pre-trained CNNs (F_t^v, F_t^a) to the region-wise (h_t^v, h_t^a) and segment-wise (s_t^v, s_t^a) attended features (as in equation 12). Unlike [51] where the coefficients come from a non-parametric correlation function, they are learnt by the network implicitly in our case. This similarity measure sim_t is concatenated with h_t^v, h_t^a, s_t^v and s_t^a to get the final aggregated features.

3.6. Supervised and Weakly Supervised event localization

The aggregated features are fed to a shared fully connected layer with softmax function to predict the probability distribution over C event categories for each input

audio-visual segment in the supervised event localization task. Multi-class cross entropy loss is used to train the network. In case of weakly supervised event localization task, it is formulated as a MIL (Multiple Instance Learning) problem [46]. Similar to [44], the predictions for each segment are aggregated to get a video-level prediction using MIL pooling, which averages over all predictions to get a single prediction for one video. During testing, event category is predicted for each segment.

3.7. Unsupervised Sound Source Localization

Event labels are not required if we want to know only the location of sound source, in a visual scene. So the goal here is to localize sound with only the video as input. This is the first work on performing sound localization using unsupervised learning on AVE dataset [44]. In order to localize sound without giving any event label, an additional negative audio sample is given to the network. That is, apart from the video and its corresponding audio (positive sample), an audio sample from a random video is considered as the negative sample. Our model is made to learn the sound localization by minimizing the distance between visual content and the positive audio sample and maximizing the distance between the visual content and the negative audio sample based on Triplet Loss.

The feature representations for visual (s_t^v), positive audio (s_t^{a+}) and negative audio (s_t^{a-}) samples from the Segment-Wise Attention Block (SWAB) are used to compute the cost. Let $\{s_t^v\}_{t=1}^{\mathcal{T}}, \{s_t^{a+}\}_{t=1}^{\mathcal{T}}$ and $\{s_t^{a-}\}_{t=1}^{\mathcal{T}}$ be denoted as $\mathbb{Z} = [s_1^v, ..., s_{\mathcal{T}}^v]^T$, $\mathbb{Y}^+ = [s_1^{a+}, ..., s_{\mathcal{T}}^{a+}]^T$ and $\mathbb{Y}^- = [s_1^{a-}, ..., s_{\mathcal{T}}^{a-}]^T$ respectively. In order to capture the inter-segment dynamic interactions across the two modalities, their corresponding Gram matrices are computed as:

$$G(\mathbb{Z}) = \begin{bmatrix} < s_1^v, s_1^v > & < s_1^v, s_2^v > & \dots & < s_1^v, s_{\mathcal{T}}^v > \\ < s_2^v, s_1^v > & < s_2^v, s_2^v > & \dots & < s_2^v, s_{\mathcal{T}}^v > \\ \vdots & \vdots & \ddots & \vdots \\ < s_{\mathcal{T}}^v, s_1^v > & < s_{\mathcal{T}}^v, s_2^v > & \dots & < s_{\mathcal{T}}^v, s_{\mathcal{T}}^v > \end{bmatrix} \tag{13}$$

where, $< s_i^v, s_j^v >$ refers to the inner product. The $\mathcal{T} \times \mathcal{T}$ Gram matrix $G(\mathbb{Z})$ captures the dynamic inter-segment interactions efficiently. $G(\mathbb{Z})$ is further normalized using the Frobenius normalization as:

$$\mathcal{G}(\mathbb{Z}) = G(\mathbb{Z})/||G(\mathbb{Z})||_F = \mathbb{Z}\mathbb{Z}^T/||\mathbb{Z}\mathbb{Z}^T||_F \tag{14}$$

$\mathcal{G}(\mathbb{Y}^+)$ and $\mathcal{G}(\mathbb{Y}^-)$ are calculated similarly. The two distance terms required to calculate the Triplet Loss are:

$$d^{pos} = < \mathcal{G}(\mathbb{Z}), \mathcal{G}(\mathbb{Y}^+) >_F ; d^{neg} = < \mathcal{G}(\mathbb{Z}), \mathcal{G}(\mathbb{Y}^-) >_F \tag{15}$$

where, $< \mathcal{G}(\mathbb{Z}), \mathcal{G}(\mathbb{M}) >_F = \sum_{i,j=1}^{\mathcal{T}} \mathcal{G}(\mathbb{Z})_{ij} \mathcal{G}(\mathbb{M})_{ij}$ refers to Frobenius inner product ($\mathbb{M}$ can be $\mathbb{Y}^+$ or $\mathbb{Y}^-$). This gives cosine similarity between the two Gram matrices. The aim is to maximize similarity between visual content and its audio counterpart and minimize the similarity between visual content and the negative audio sample. This is done using the proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function, as:

$$L_{unsup} = \frac{1}{N} \sum_{i=1}^{N} max(d_i^{neg} - d_i^{pos} + \gamma, 0) \tag{16}$$

where, γ is the margin which is empirically chosen as 0.7 and N is the number of training samples.

4. Experiments and Results

4.1. Dataset used

The *Audio-Visual Event* dataset from [44] is used to evaluate our proposed method. This dataset which is a subset of Audioset [14] contains 4143 videos each of length 10 seconds, across 28 different categories. However, the duration of the events in these videos span from a minimum of 2 seconds to a maximum of 10 seconds. The dataset encompasses a wide and diverse range of event categories like human speeches, animal sounds, musical performances, vehicle sounds, *etc*. Labels are available video-wise as well as segment-wise with clearly demarcated temporal boundaries. Each category contains 60 to 188 video shots.

4.2. Implementation Details

The audio and visual features from raw audio and visual segments are extracted using pre-trained CNNs as in [44]. The visual features are extracted using VGG-19 [40] pre-trained on ImageNet [35] for every one second segment in the video while the audio features are extracted using [17] which bears a close resemblance to VGG architecture. The VGG architectures are used to ensure a fair comparison with the existing state-of-the-art methods.

Method	Sup. Acc.	W-Sup. Acc.
Audio	60.6	57.9
Visual	56.7	54.3
Audio-assisted Visual	59.9	56.5
Audio + Visual	73.1	65.8
AVE [44]	72.7	66.7
AVSDN [24]	72.8	66.5
Ours (Aud + Aud-ass. Vis)	**74.8**	**68.9**

Table 1. **Performance comparison** (in %) of various methods employed for supervised and weakly supervised event localization tasks. For a fair comparison, we use our implementation version of AVSDN [24], which uses VGG-19 to extract visual features.

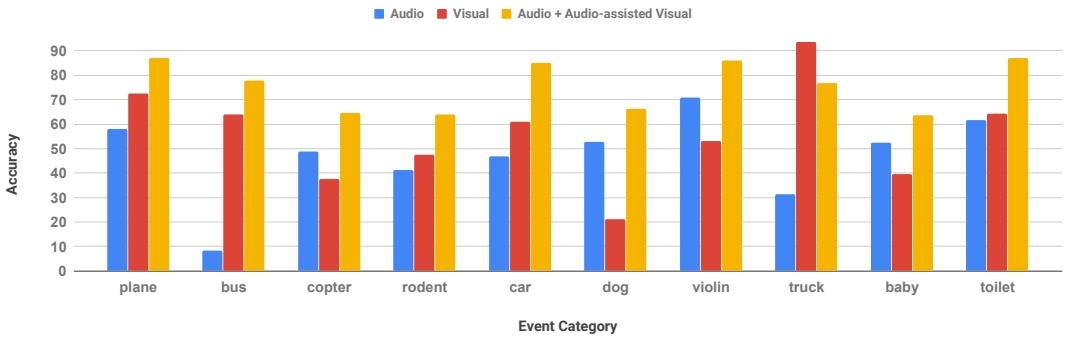

Figure 5. Bar chart depicting accuracies of a few selected event categories, as obtained using 3 variants of our proposed model (supervised) that use: only audio (A), only visual (V), audio-assisted visual plus audio features respectively.

4.3. Performance Analysis

The performance of our proposed method is evaluated against various other models (as shown in table 1) for both supervised and weakly supervised tasks. The final aggregation of audio and visual features is effective in spatial localization of sound in video. Aggregating audio-assisted visual features with audio features gives the best accuracy in our model (last row of table 1). Our proposed model is also evaluated against other existing works [24, 44] and our model beats them by a significant margin. Also, it can be observed that using only audio features outperforms the case of using only visual features. Since audio-assisted visual features contain some amount of audio information, they are more accurate than using plain visual features.

4.4. Ablation Studies

Different Fusion Techniques Interactions between the two modalities determine the performance of event localization (as is evident from table 1). While simple element-wise operations can also be used for fusion, it was empirically observed that they fail to absorb high-level associations between the two modalities. This is shown in table 2 which analyzes the performance of different fusion strategies. Also, the best performance is achieved using our AVFB model compared to that obtained using LSTM-based fusion and MFB [50] separately. This shows that a combination of neural network based fusion and bilinear pooling based fusion outperforms other fusion techniques.

Analyzing accuracies of individual event categories Figure 5 shows the accuracies for a few event categories (the rest shown in Supplementary) in groups of three bar plots, as obtained by our proposed model, that uses only audio, only visual and audio-assisted visual features with audio features respectively. The fusion gives an obvious improvement in accuracy in most cases compared to that using only the features from the two modalities without fusion.

Importance of each module in the architecture: Table 3 shows the significance of the three modules: AVFB, SWAB and Aggregation in our architecture. Using mean and max pooling together in SWAB gives a minor improvement in performance than using them separately (as shown in table 3). The final step of aggregation of the two feature representations obtained from the two modalities also leads to substantial improvement in performance. On the whole, all three modules play a vital role in capturing high-level associations between the two modalities resulting in getting compact feature representations for sound localization task.

Fusion strategy	Accuracy (in %)
Element-wise multiplication	67.7
Element-wise addition	69.3
Concatenation + FC	68.8
LSTM	71.1
MFB [50]	73.2
AVFB (LSTM + MFB)*	**74.8**

Table 2. **Different Fusion Strategies.** Accuracies of models employing different audio-visual fusion techniques (in supervised learning setup) are compared against our AVFB. Rest of the model (SWAB and aggregation) is kept intact. Concatenation + FC refers to fusing by concatenating features followed by passing it to a fully connected layer. $*$ - identical to the last row of table 1.

Model	Accuracy (in %)
AVFB	73.4
AVFB + SWAB (only mean)	74.0
AVFB + SWAB (only max)	73.7
AVFB + SWAB (mean + max)	74.4
AVFB + SWAB + Aggregation	**74.8**

Table 3. **Ablation Study.** The importance of each of the three modules: AVFB, SWAB and Aggregation is demonstrated (in supervised learning case). A combination of all the three modules gives the best accuracy.

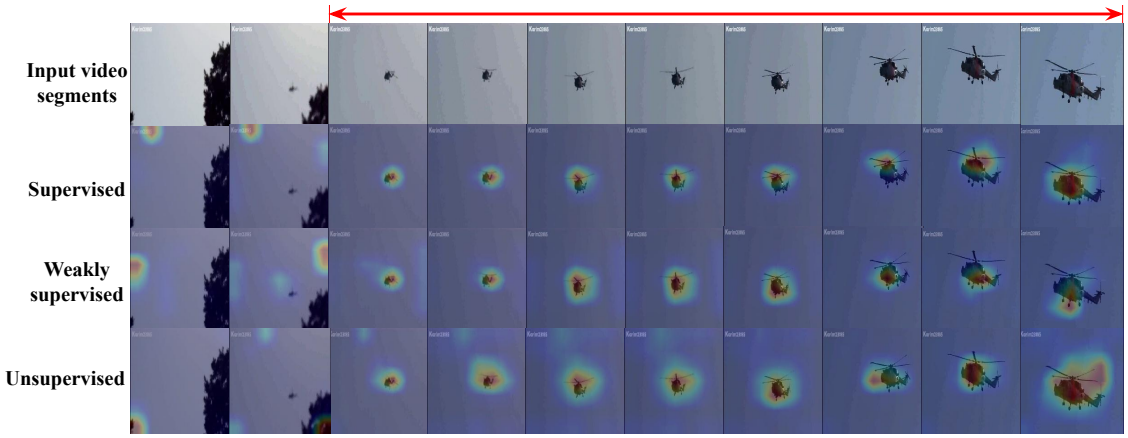

Figure 6. **Attention Maps** in a video obtained from supervised, weakly supervised and unsupervised models, for each of the 10 segments. The arrow on top of the figure indicates the presence of event (ground truth) in the video, where the sound is audible as well as the object producing the sound is visible. Since the first two segments do not exhibit the event, attention maps randomly map to the background.

4.5. Qualitative Results

Figures 6, 7 and 8 show some qualitative results of our model. Additional visual results are shown in the Supplementary section. The attention maps obtained using supervised, weakly supervised and unsupervised models are shown in figure 6. It can be seen that the attention maps obtained using unsupervised learning are almost as precise compared to supervised and weakly supervised tasks, even without the use of event labels. Figure 7 shows the attention maps obtained by unsupervised sound localization for various event categories. It can be seen that the model works well across a diverse set of events. Figure 8 leads to an important observation that the attention gets mapped based on the audio given and not just based on the salient objects present in the scene thus indicating the importance of audio-visual fusion and aggregation. This gives an affirmation that the model relies majorly on audio for decision making.

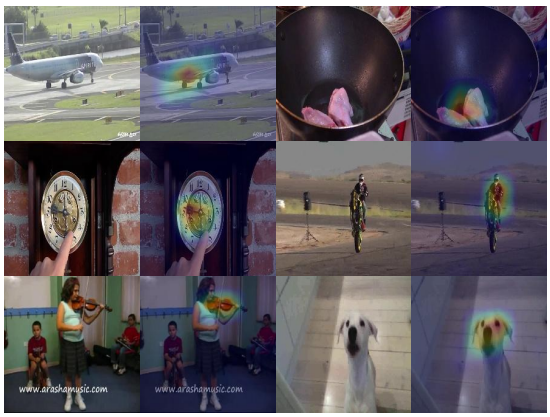

Figure 7. Attention maps obtained using unsupervised learning across various event categories. Input frames along with the learnt attention maps are shown.

Figure 8. Input frames along with the attention maps for two different events showing that the attention is learnt based on the audio and visual features and not based on the salient objects present.

5. Discussion and Conclusion

This paper proposes a method with two novel blocks: Audio Visual Fusion Block (AVFB) and Segment-Wise Attention Block (SWAB) to tackle sound source localization using supervised and weakly supervised learning. We also propose a novel Audio Visual Triplet Gram Matrix Loss (AVTGML) function to localize sound source using unsupervised learning. We demonstrate empirically that jointly modeling audio and visual content captures high-level semantic information leading to better performance. We show the importance of our proposed Audio Visual Fusion Block (AVFB), Segment-Wise Attention Block (SWAB) and the aggregation block, through extensive experiments. We have also demonstrated that the sound source localization performed using unsupervised learning yields attention maps similar to that of the supervised setup. This can be attributed to the proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function which succeeds in capturing inter-segment dynamic relationships in videos. The proposed model allows the flexibility and scope of perceiving visual scenes using sound localization by a machine. This area of research further paves way to a wide range of applications like audio-visual scene understanding and cross-modal localization.

References

[1] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7558–7567, 2019. 3

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 1, 2, 3

[3] R. Arandjelovic and A. Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 2, 3

[4] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016. 2, 3

[5] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 3

[6] Z. Barzelay and Y. Y. Schechner. Harmony in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 2, 3

[7] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620, 2017. 2

[8] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, pages 251–263, 2016. 3

[9] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems*, pages 772–778, 2001. 2, 3

[10] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2

[11] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[12] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[13] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2, 6

[15] D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016. 3

[16] J. R. Hershey and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems*, pages 813–819, 2000. 2, 3

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017. 4, 6

[18] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356, 2019. 3

[19] D. Hu, F. Nie, and X. Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 3

[20] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2012. 2, 3

[21] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 88–95, 2005. 2, 3

[22] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 2

[23] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2

[24] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006, 2019. 2, 3, 6, 7

[25] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid. Visual question answering with memory-augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6975–6984, 2018. 2, 4

[26] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. 2, 4

[27] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976. 2

[28] K. Omata and K. Mogi. Fusion and combination in audio-visual integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2090):319–340, 2007. 1, 2

[29] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3

[30] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016. 3

[31] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 3

[32] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961, 2015. 2, 4

[33] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Advances in Neural Information Processing Systems*, page 5, 2015. 2, 4

[34] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba. Self-supervised audio-visual co-segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361, 2019. 2

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[36] I. Schwartz, A. G. Schwing, and T. Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12548–12558, 2019. 3

[37] J.-L. Schwartz, F. Berthommier, and C. Savariaux. Audio-visual scene analysis: evidence for a" very-early" integration process in audio-visual speech perception. In *Seventh International Conference on Spoken Language Processing*, 2002. 1, 2

[38] R. Sekular, A. Sekular, and R. Lau. Sound alters visual motion perception. *Nature*, 1997. 2

[39] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 2, 3

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[41] J. B. Tenenbaum and W. T. Freeman. Separating style and content. In *Advances in Neural Information Processing Systems*, pages 662–668, 1997. 2

[42] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 2

[43] Y. Tian, C. Guan, J. Goodman, M. Moore, and C. Xu. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872*, 2018. 3

[44] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2, 3, 4, 6, 7

[45] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5

[46] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, 2015. 6

[47] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[48] X. Xu, B. Dai, and D. Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[49] H. Xue, Z. Zhao, and D. Cai. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing (TIP)*, 26(12):5656–5666, 2017. 2, 4

[50] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1821–1830, 2017. 2, 4, 7

[51] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing (TIP)*, 2019. 5

[52] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[53] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 2, 3