Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization

Yan-Bo Lin[1] and Yu-Chiang Frank Wang[1,2]

[1] Graduate Inst. Communication Engineering, National Taiwan University, Taiwan
[2] ASUS Intelligent Cloud Services, Taiwan
{yblin98,ycwang}@ntu.edu.tw

Abstract. Audio-visual event localization requires one to identify the event label across video frames by jointly observing visual and audio information. To address this task, we propose a deep learning framework of cross-modality co-attention for video event localization. Our proposed audiovisual transformer (AV-transformer) is able to exploit intra and inter-frame visual information, with audio features jointly observed to perform co-attention over the above three modalities. With visual, temporal, and audio information observed across consecutive video frames, our model achieves promising capability in extracting informative spatial/temporal features for improved event localization. Moreover, our model is able to produce instance-level attention, which would identify image regions at the instance level which are associated with the sound/event of interest. Experiments on a benchmark dataset confirm the effectiveness of our proposed framework, with ablation studies performed to verify the design of our propose network model.

1 Introduction

In real-world activities, visual and audio signals are both perceived by humans for perceptual understanding. In other words, both visual and audio data should be jointly exploited for understanding the observed content or semantic information. Recently, audio-visual event localization [1,2,3,4,5] attracts the attention from computer vision and machine learning communities. As depicted in Fig. 1, this task requires one to identify the content information (e.g., categorical labels) for each frame or segment in an video, by observing both visual and audio features across video frames. Audio-visual event localization can be viewed as a cross-modality learning task, which deals with the challenging task that the feature representations and distributions across visual and audio domains are very different. To explore audiovisual representation, joint learning of multi-modal deep networks across these two domains have been studied, e.g., classification [6,7], lip reading [8,9,10] and sound synthesis [11,12]. These works demonstrate that such audio-visual based models can be applied to several downstream tasks. However, these models rely on the simultaneous presence of both visual and audio information. In other words, they cannot deal with scenarios with partial modality information observed. On the other hand, methods for locating sound

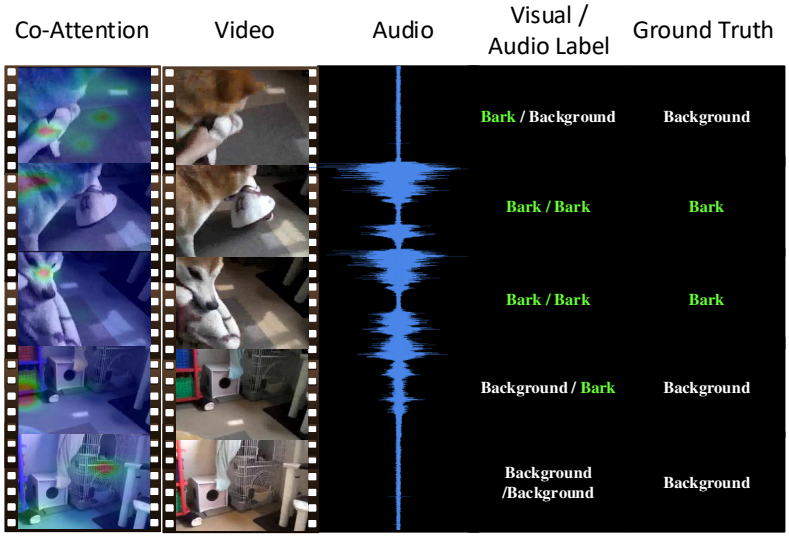

Fig. 1: **Illustration of audio-visual event localization** (recognizing video event with matched visual and audio information). Note that the first column shows our correct localization outputs with cross-modality co-attention, the 2nd and 3rd columns show the video and audio inputs across five consecutive frames, with ground truth visual/audio and event labels depicted in the last two columns.

source models [13,14,15,16] have also been proposed via exploiting mutual information between audio and visual data. However, these models cannot easily attend proper regions of instances or distinguish sounding objects from silent ones.

To this end, we propose a novel deep attention model which jointly performs visual, temporal, and audio cross-modality co-attention to better associate audio and visual information for video event localization. This is realized by our proposed audiovisual transformer (AV-transformer) for jointly encoding intra-frame and inter-frame patches, followed by exploitation of encoded intra-frame and inter-frame visual and audio features. As a result, one important features of our proposed attention model is that we not only improve the overall localization (i.e., classification) performances, it further attends proper regions across video frames (e.g., the corresponding object of interest in Fig. 1). More importantly, we will show that our model is not limited to the use of fully supervised video data (i.e., visual and audio labels annotated for each frame). Learning of our model in a weakly supervised setting can be conducted, in which only an overall soft label at the video level is observed during training.

The contributions of this work are highlight below:

- We propose an audiovisual transformer (AV-transformer) for visual, temporal, and audio co-attention, with the goal of solving event localization tasks.
- Without attention supervision during training, our model is able to perform instance-level attention by jointly encoding intra and inter-frame image patches and audio features.
- Experimental results demonstrate that our proposed model performs favorably against state-of-the-art approaches in various settings, while instance-level attention can be additionally achieved by our model.

2 Related Work

Video Classification. Methods based on deep neural networks have shown promising performances on the task of video classification [17,18,19,20,21,22,23], which takes visual and temporal information for predicting action or event categories for input videos. To explore the aforementioned spatial-temporal features from videos, 3D convolutional networks are utilized, in which 3D architectures with 3D kernels are considered [24,25,26]. On the other hand, long short term memory (LSTM) networks [17] has also been employed to observe 2D CNN features over time. Such recurrent neural networks (RNN) [17,27,28] are alternative ways to learn the temporal relation between frames. However, since uses of RNNs might limit the length of the input video to be observed [27,28], some works choose to sample frames from the entire video to learn robust reasoning relational representation [23,20,21,29,30].

Relating Audio and Visual Features. While RNN-based models have been widely applied to extract spatial-temporal features from videos, such methods do not consider audio features when modeling temporal information. To address this issue, cross-modality learning using audio and visual data are proposed [31,32,33,12,34,35]. For example, Aytar et al. [32] learn the joint representation from audio-visual data, with the goal to identifying the content using data in either modality. Arandjelovic and Zisserman [31] also exploit the variety of audio-visual information for learning better representation in audio-visual correspondence tasks. Furthermore, they [15] visualize sound localization in visual scenes, which would serve as the bridge connecting between audio and visual modality. Owens et al. [16] leverage ambient sounds when observing visual contents to learn robust audio-visual representations. The resulting representation is further utilized to perform video tasks of action recognition, visualization of the sound sources, and on/off-screen source separation. These studies [16,15] apparently show that sound source localization can be guided by semantic visual-spatial information, and verify that these cross-modality features would be beneficial in the aforementioned video-based applications.

Aside from learning audio-visual representation, works like [8,9,10] demonstrate that such audio-visual based models can be applied to synthesize videos with face images (e.g., with lip motion), corresponding to the input free-form spoken audio. Concurrently, some audio-related tasks [16,13,36,34,37] also utilize visual representation to solve speech separation [37,16], musical instruments

[13,34,38,39,40] and objects [36]. Most of these methods maintain a "mix and separate" training scenario where the training videos are first mixed and separated afterward. For instance, MP-Net [39] initially separates sound with large energy from mixtures which is composed of any arbitrary number of sounds. the sound with small magnitude would keep emerging afterwards. Furthermore, by detecting sounding objects to guide the learning process using unlabeled video data, Gao et al. [40] propose a framework to bridge the localized object regions in a video with the corresponding sounds to achieve instance-level audio-visual source separation. Methods like [38,41] utilize visual motions or body gesture to separate sound signals, and thus audio-visual source separation can be performed for different instruments.

Nevertheless, while the aforementioned works show promising results in learning audio-visual representation, it is still challenging to address audio-visual event localization, which requires one to identify the event with both visual and audio modality properly presented, especially in a weakly supervised setting (i.e., no frame-level ground truth annotation).

Audio-visual Event Localization. Audio visual event localization aims to detect events in videos, which requires both audio and visual activities and events to be identified. Early works [1,42,3,4] are proposed to jointly learn audio-visual information in each local segments of the input video. However, due to potential inconsistency between information observed from audio and visual signals, the data from either modality with insignificant cues may interfere the event prediction. Therefore, works [2,5] tackle with this issue by disregarding information from audio/visual data with irrelevant categorical events. Nevertheless, the aforementioned methods only consider the correlation between audio and visual in a video segment at the same time frame. To address this issue, we further jointly exploit relationship between visual patches and audio signal from video segments within the same or across time interval, which allows our model to learn segment-wise events with the guidance of nearby audio and visual frames. In the next section, we will present and discuss the details of our proposed co-attention model, which jointly exploits visual, temporal, and audio data for improved localization and instance-level visualization.

3 Proposed Method

3.1 Notations and Problem Formulation

In this paper, we design a novel deep neural network model for audio-visual event localization. In order to deal with cross-modality signals observed from audio and video data with the ability to identify the event of interest, our model exploits visual information within and across video frames. Together with the audio tracks, the proposed model not only performs satisfactory localization performances, it also exhibits promising capability in attending the objects in the input video associated to that event.

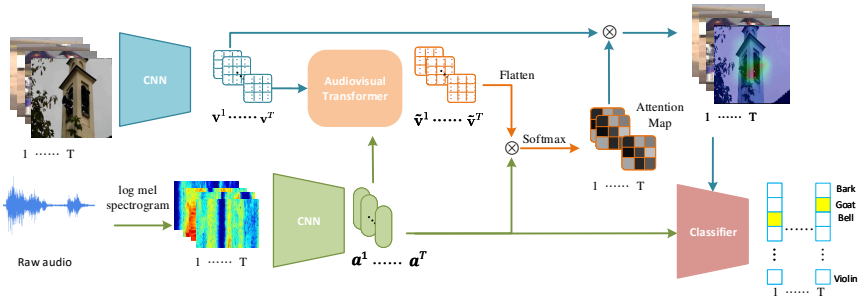

Fig. 2: **Overview of our proposed audiovisual transformer (AV-transformer) for instance-attended audio-visual event localization**. Note that $\otimes$ denotes matrix multiplication, and the softmax operation is performed on each row in the AV-transfomer module.

For the sake of completeness, we first define the settings and notations which will be used in this paper. Following [1], two training schemes for audio-visual event localization are considered: *supervised* and *weakly-supervised* learning. Given a video sequence with T seconds long, it is split audio a and video v tracks separately into T non-overlapping segments $\{a^t, v^t\}_{t=1}^{T}$, where each segment is 1s long (since the event boundary is labeled at second-level). For the supervised setting, segment-wised labels denote $\mathbf{y}^t = \left\{ y_k^t | y_k^t \in \{0,1\}, \sum_{k=1}^{C+1} y_k^t = 1, t \in \mathbb{N} \right\}$, $\mathbf{y}^t \in \mathbb{R}^{C+1}$, where t denotes the segment index and C denotes total event categories. We note that, considering the category of background, the total number of event categories becomes $C+1$. In the supervised setting, every segment-wise labels are observed during the training phase.

As for the scheme of weakly-supervised learning, we only access to the video-level event labels $\mathbf{Y} \in \mathbb{R}^{C+1}$ during the training phase (e.g., event category for a whole video). As for the background event, we take different event categories from audio and visual contents as inputs (e.g., dog image and goat sound). Note that predicted video-level event labels are processed by max pooling through time the segment event labels $\hat{\mathbf{m}} = \max\{\mathbf{m}^t\}_{t=1}^{T}$, where $\mathbf{m}, \hat{\mathbf{m}} \in \mathbb{R}^{C+1}$. $\mathbf{m}^t$ is the prediction from audio-visual event localization network. For this weakly supervised setting, while it is less likely to be affected by noise from either modality at the segment level during training, it also makes the learning of our model more difficult.

Fig 2 depicts our proposed AV-transformer for audio-visual event localization. It is worth noting that, cross-modality instance-level attention can be performed by our proposed framework. We now discuss the details of our model in the following subsections.

3.2 Jointly Learning Intra and Inter-Frame Visual Representation

Visual attention has been widely utilized in recent VQA and audio-visual related tasks [43,44,45,46,47,1,14,15]. Although convolution neural networks have been successfully applied in the above works to identify spatial regions of interest with impressive results, such attention is typically performed at the pixel level, based on the information observed for the corresponding tasks (e.g., guidance at the network outputs) [14,15,13,16,47].

For the task of audio-visual event localization, one needs to identify the video segments with the event of interest. It would be preferable if one can attend on the object of interest at the instance level during localization, which would further improve the localization accuracy.

Previously, [48] considered local context information by feeding image patches into a LSTM, which can facilitate understanding objects in image scenes for VQA tasks. [49] introduced non local block for video understanding. These blocks are used to learn visual context information from image patches through space, time or spacetime by transformer encoder [50]. Thus, inspired by [48] and [49], we present a unique audiovisual transformer (AV-transformer) as shown in Fig 3 to encode local context information into proper representation, so that object instances corresponding to event of interest can be attended accordingly. To achieve this goal, we input local image patches of successive video frames and audio segments into our audiovisual transformer, which encodes the image patches of that frame in a sequential yet temporal visual features.

More precisely, we divide a input video frame at time step t into R patches, and extract the CNN feature for each patch. These visual representations of each region are denoted as $\{\mathbf{v}_r^t, r = 1, 2..., R\}$, where $\mathbf{v}_r^t \in \mathbb{R}^{1\times K}$ represents the visual features of the rth patch. These visual features and audio features $\mathbf{a}^t$ are served as the inputs to the audiovisual transformer, which is described below:

$$\tilde{\mathbf{v}}^t = Trans(\mathbf{v}^{t-1}, \mathbf{v}^t, \mathbf{v}^{t+1}, \mathbf{a}^t), \tag{1}$$

where $Trans(.)$ denotes the audiovisual transformer, the attention block in our unique audiovisual transformer can be further described as follows:

$$\begin{aligned} \mathbf{Attn}_{r,i}^t &= (\theta(\mathbf{v}_r^t) + \mathbf{a}^t) \sum_{k=t-1}^{t+1} \phi(\mathbf{v}_i^k) \quad \forall i \in R, \\ \tilde{\mathbf{v}}_r^t &= \sum_{\forall r,i\in R} \text{Softmax}(\mathbf{Attn}_{r,i}^t)(\mathbf{v}_r^t) \end{aligned} \tag{2}$$

where $\tilde{\mathbf{v}}_r^t$ indicates intra/inter-frame visual representation of the rth image patch at audio instant t. **Attn** shows the relation between patch i and j at time t. $\theta(.)$ and $\phi(.)$ are multilayer perceptrons. We note that the Softmax(.) operation is performed at each row. We gather R patches for co-attention visual representations of video frame at time t, that is, $\tilde{\mathbf{v}}^t = \{\tilde{\mathbf{v}}_1^t, ..., \tilde{\mathbf{v}}_R^t\} \in \mathbb{R}^{R\times K}$.

It can be seen that, by advancing our audiovisual transformer, visual representation encoded would describe local spatial and temporal information within

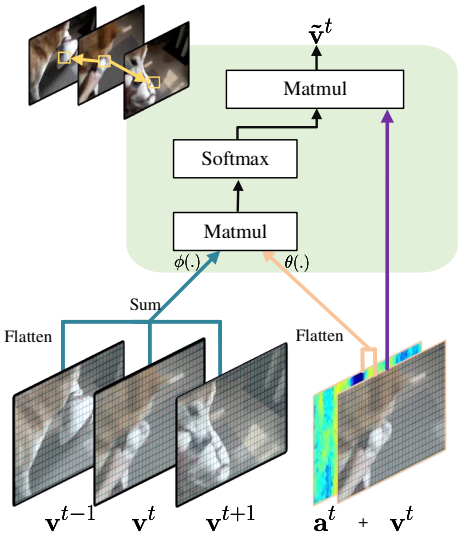

Fig. 3: **Audiovisual transformer**: visual feature $\tilde{\mathbf{v}}^t$ is encoded by jointly observing audio $\mathbf{a}^t$ and $\mathbf{v}^t$ visual context at time t, followed by visual context over consecutive frames ($\mathbf{v}^{t-1}$, $\mathbf{v}^t$ and $\mathbf{v}^{t+1}$).

successive video frames. By combing local, temporal and audio information in this stages, this audiovisual transformer allows improved attention at the instance-level as later verified.

3.3 Instance Attention for Event Localization

The visual encoders introduced in the previous subsections exploit local spatial and short-term temporal information. As noted above, to perform frame-level audio-visual event localization, it would be necessary to integrate the audio features into consideration.

Some previous works [14,15,13,16] have presented to explore the relationship between audio and visual scenes. They show that correlations between these two modalities can be utilized to find image regions that are highly correlated to the audio signal. However, these works only consider single image inputs and its corresponding sound signals, which might result in incorrect association due to overfitting to the visual content. Another concern is that, if more than one instance visually correspond to the event of interest, how to identify the object instance would not be a trivial task. Take an audio-visual event in which a person is playing violin solo in a string quartet for example, it would be challenging to identify which image region is related to the audio signal, if only a single frame input is observed.

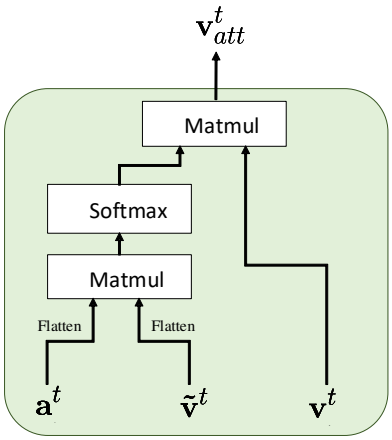

Fig. 4: **Instance Attention**: Observing locally and temporally visual-attended features $\tilde{\mathbf{v}}^t$ and audio inputs $\mathbf{a}^t$ to output the final co-attention features $\mathbf{v}_{att}^t$ for audiovisual event localization. Note that we remove the subscript r (patch index) of visual features for simplicity.

To address the above challenge, we propose to perform cross-modality co-attention over visual, temporal, and audio features. By taking temporal information into consideration, our intra and inter-frame visual features would be associated with the audio features, which would make the localization of audio-visual events more applicable. To achieve this goal, we advance the concept of self-attention [50] for computing a soft confidence score map, indicating the correlation between the attended visual and audio features. Different from existing co-attention mechanisms like [1,14,15,13,48,51], our input visual features jointly take spatial and temporal information via intra and inter-frame encoding, followed by joint attention of audio features. Thus, our co-attention model would be more robust due to the joint consideration of information observed from three distinct yet relevant data modalities.

As depicted in Fig. 4, we obtain the r th local visual feature $\tilde{\mathbf{v}}_r^t$ at time t, where $r = 1, ..., R$, and our co-attention model aims to produce the weight to depict how relevant $\mathbf{v}_r^t$ and $\mathbf{a}^t$ is. The attention score M_r^t can be interpreted as the probability that location r is the right location related to the sound context. Note that M_r^t in our co-attention model is computed by:

$$M_r^t = \text{Softmax}(\tilde{\mathbf{v}}_r^t \cdot (\mathbf{a}^t)^{'}), \tag{3}$$

where $\cdot$ indicates the dot product and $'$ denotes transpose operation. Note that visual and audio representation are in the same dimension, that is, $\tilde{\mathbf{v}}_r^t, \mathbf{a}^t \in \mathbb{R}^{1\times D}$. With all local visual features are observed, we pool the associated outputs by a weighted sum M to obtain the final visual attention representation of the

image at time t, i.e.,

$$\mathbf{v}_{att}^{t} = \sum_{r=1}^{R} M_r^t \mathbf{v}_r^t. \tag{4}$$

With this cross-modality co-attention mechanism, our visual attention feature $\mathbf{v}_{att}^{t}$ would exclude local image regions which are irrelevant to the audio signal, and better bridges between the visual content and the audio concept by preserving the audio-related image regions. This is the reason why *instance-level* visual attention can be performed. We note that, this attention feature $\mathbf{v}_{att}^{t}$ can be easily deployed in current event localization models (e.g., [1,42]). We will detail this implementation and provide thorough comparisons in the experiment section.

4 Experiments

4.1 Dataset

For the audio-visual event localization, we follow [1] and consider the *Audio-Visual Event* (AVE) [1] dataset (a subset of Audioset [52]) for experiments (e.g., Church bell, Dog barking, Truck, Bus, Clock, Violin, etc.). This AVE dataset includes 4143 videos with 28 categories, and audio-visual labels are annotated at every second.

4.2 Implementation Details

In this section, we present the implementation details about the evaluation frameworks. For visual embedding, we utilize the VGG-19 [53] pre-trained on ImageNet [54] to extract 512-dimensional visual feature for each frame. The feature map of whole video frames with T seconds long is $\mathbb{R}^{T\times 7\times 7\times 512}$. We obtain 7×7 channels and 512 dimension with each channel. Each channel is processed by multilayer perceptrons (MLP) into 128 dimensions. Then, we reshape 7×7 channels into 49 channels corresponding to aforementioned total image regions R. As for audio embedding, we extract a 128-dimensional audio representation for each 1-second audio segment via VGGish [55], which is pre-trained on YouTube-8M [52]. Thus, we have audio features produced in a total of T seconds, i.e., $\mathbb{R}^{T\times 128}$.

For both fully supervised and weakly-supervised audio-visual event localization, we consider **frame-wise accuracy** as the evaluation metric. That is, we compute the percentage of correct matchings over all test input frames as the prediction accuracy. We note that, for fair comparisons, we apply VGG-19 as the visual backbone and VGGish as audio embedding models.

4.3 Experiment results

Quantitative results. We compare the performance of supervised event localization using baseline and recent models. [1] choose to concatenate audio and

Table 1: Performance comparisons using baseline or state-of-the-art localization methods in **supervised** (i.e., ground truth y_t observed for each frame during training) and **weakly supervised** (i.e., only ground truth $\mathbf{Y}$ observed for training). The numbers in bold indicate the best results (i.e., methods with our proposed instance attention mechanism). $*$ indicates the reproduced performance using the same pre-trained VGG-19 feature and the same weakly supervised setting for a fair comparison.

Method	Accuracy (%) Fully Supervised	Accuracy (%) Weakly Supervised
Audio only+LSTM	59.5	53.4
Visual only+LSTM	55.3	52.9
AVEL [1]	71.4	63.7
AVSDN [42]	72.6	68.4
AVEL+Att [1]	72.7	66.7
DAM [2]	74.5	-
Xuan et al.*[5]	75.1	67.8
Ramaswamy et al.[4]	74.8	68.9
AVIN [3]	75.2	69.4
AVSDN+Ours	75.8	**70.2**
AVEL+Ours	**76.8**	68.9

visual outputs from LSTMs and audio-guided visual attention. [42] apply an additional LSTM to serve as the final prediction classifier. DAM [2] advance state-of-the-art results in this task by jointly exploiting audiovisual relevant events. Note that, DAM requires event labels at each segment for calculating the audiovisual segment relevance. Thus, DAM would not be evaluated in the weakly supervised setting. In this work, we adapt our instance attention visual features in AVEL and AVSDN. Table 1 summarizes the performances of our methods and others in fully supervised on the AVE dataset. As for the weakly supervised setting, we repeat the same experiments and list the performance comparisons also in Table 1.

From tables presented, it is clear that use of our instance attention features would increase the localization accuracy. In other words, either observing frame-level or video-level labels, our proposed audiovisual transformer would properly extract cross-modality features for improved audio-visual event localization.

Qualitative results. We now present example visualization results in fully supervised settings using the AVEL classifier. The attention output produced by ours and the method of [1] are shown in Fig. 5 and 6, respectively. We note that, the current co-attention method [1] (Fig. 5) only considers the patch-wise relationship between audio notion and visual feature extracted by CNN. Thus, this mechanism only focuses on local regions but ignores relationship between neighboring patches. To address this issue, our AV-transformer jointly exploits intra-frame and inter-frame visual patches and audio information. The

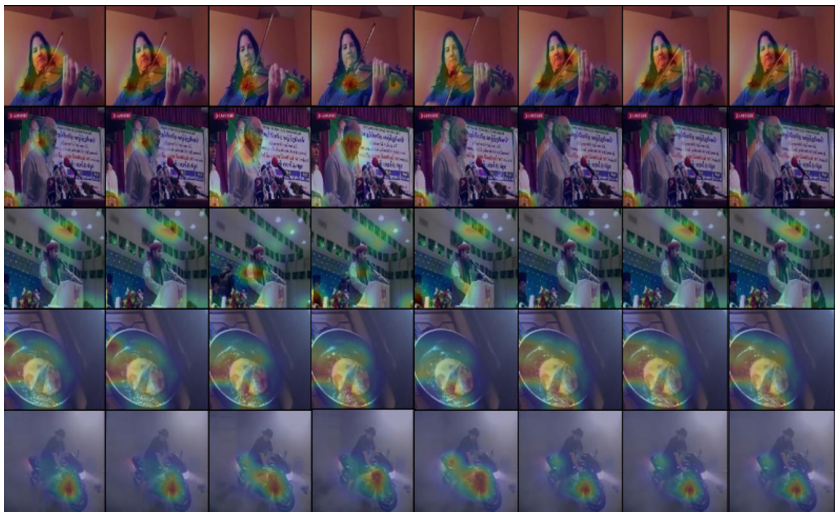

Fig. 5: **Example attention results using AVEL [1] with their attention model of AVEL [1]:** Each row shows a video input with visually attended regions. Take row 2 and 3 for example, it can be seen that AVEL with their attention model would incorrectly attend the regions of human voice which was not actually associated with the sound of human voice.

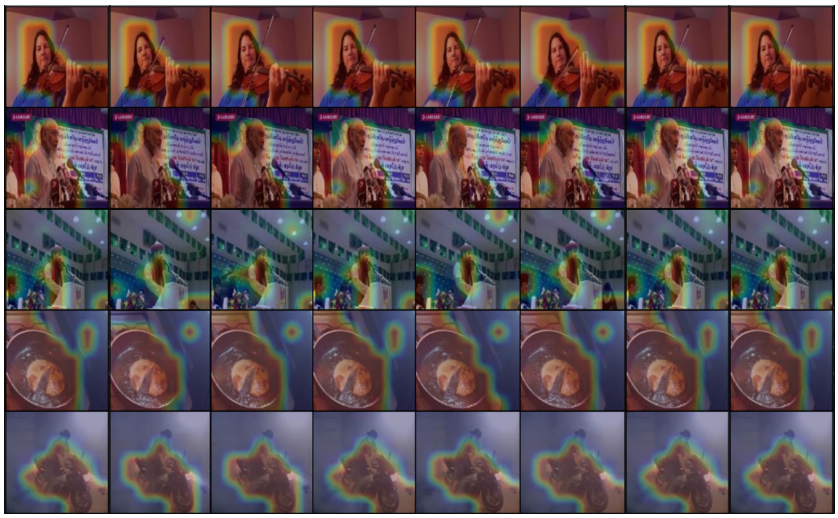

Fig. 6: **Example attention results using AVEL [1] with our instance attention model:** Each row shows a video input with visually attended regions. It can be seen that our model produced satisfactory attention outputs with the corresponding audio-visual events.

encoded visual features derived from our AV-transformer preserve not only local patches information but neighboring patches with the same semantics. Furthermore, the inter-frame representation can facilitate the smoothness of attended regions across frames. We note that, in the second and third rows, there were several non-interested people in the background. It would distracted the attention. Our method could attend on the proper regions under the similar object in the scene. In the forth and fifth rows, there were people cooking and riding motorcycle. Our model is able to not only attend on sounding objects but precisely preserve edge of objects.

The above quantitative and qualitative result successfully verify the effectiveness and robustness of our proposed cross-modality co-attention model. It not only produces improved audio-visual event localization result; more importantly, it is able to attend visually informative local regions across frames, and performs instance-level visual attention. This is also the reason why improved event localization performances can be expected.

User studies. To evaluate the quality of qualitative results , we invite 20 people to watch the video with different attention results from audio-guided [1], audio-visual object localization [15,14], and ours. The participants voted the best results of three samples for each instance. We observe 72.2% (Ours), 16.7% ([1]), and 11.1% ([15,14]). These results would also support our quantitative comparisons in Table 1.

4.4 Ablation studies

In this section, we verify the design and contributions of our audiovisual transformer and different visual co-attention mechanisms [1,15,14]. This would support the learning and exploitation of intra and inter-frame visual representation for audio-visual event localization.

Inter-frame visual representation. As to study the effects of learning inter-frame visual representations for instance attention, we consider different methods to model such inter-frame visual features. To model across frames visual representation, we utilize 3D convolutional networks [24] (Conv3D) and LSTM [56] network in our work. We note that, for standard Convolutional Neural Network [57] and the recent I3D Network [26], both based on consecutive video frames and optical flow, are also able to perform such modeling. In this ablation study, for fair comparisons, we only consider Conv3D and LSTM which do not require calculation of optical flow information. As for Conv3D, the inter-frame visual features can be modeled by Conv3D directly. However, LSTM only receives 1D embedding over times. Thus, we use the same location at every video frame as 1D embedding vector sequence, then the LSTM is applied to model temporal feature until every location across frames are processed.

We note that, the visual features derived from multilayer perceptron (MLP), Conv3D, LSTM and our AV-transformer are able to be utilized in current co-attention [1,14,15] methods. There are two typical co-attention mechanisms: audio-guided (AG) [1] and audio-visual object localization (AVOL) [15,14] co-attention. To be more specific, AVOL measures the correlation between visual

Table 2: Comparisons of recent audio-visual co-attention mechanisms [1,15,14] with integrating different visual representation in fully **supervised** setting (i.e., all ground truth y_t observed during training). The numbers in bold indicate the best results (i.e., with our instance attention).

Visual Representation	Classifier			
	MLP	LSTM	AVSDN	AVEL
MLP+AG	64.0	69.0	73.1	72.7
MLP+AVOL	65.2	70.1	73.7	74.8
Conv3d+AG	64.9	71.0	73.8	75.2
Conv3d+AVOL	64.5	69.4	73.6	73.6
LSTM+AG	61.2	67.7	72.6	74.3
LSTM+AVOL	66.9	70.5	73.2	75.4
AV-Transformer+AG	65.9	67.2	74.0	74.1
Ours	**67.6**	**71.4**	**75.8**	**76.8**

patches and audio data based on cosine similarity, while AG determines the associated correlation via learning a neural net. Note that, we use AVOL in our instance attention. Therefore, we not only present different methods to encode inter-frame visual features but also test them on the two co-attention methods. As shown in Table 2, our instance attention performs favorably against other models with inter-frame visual encoding. In this table, the suffix of visual representation is the co-attention method (e.g., AG and AVOL). It is also worth noting that, our method also performed against different co-attention mechanisms. Another advantage of our approach is that, since our inter-frame visual features are calculated by intra-frames regardless the fixed kernel size, whose computation cost is lower than the models using Conv3D and LSTM. Based on the above results and observations, we can also confirm the jointly learning of intra and inter-frame visual features would be preferable in our cross-modality co-attention model, which would result in satisfactory event localization performances.

Audiovisual representation of AV-transformer. As to study the effects of jointly learning intra-frame and inter-frame visual representations in our audio-visual transformer. Besides, we exploit the audio interaction between each visual patch. As shown in Table 3, intra shows the only usage of time t frame in our AV-transformer, which means only $k = t$ in Eq. (2) without audio feature $\mathbf{a}^t$. Inter indicates three successive time step $k = t-1, t, t+1$, and audio illustrates audio feature $\mathbf{a}^t$ in Eq. (2). Table 3 verifies the effectiveness of jointly considering spatial, temporal visual information and audio signals for cross-modality co-attention for audio visual event localization to better associate audio and visual information for the task.

Table 3: Comparisons of audiovisual representation in our AV-transformer. We explore different interactions on across frames, single frames and audio in fully **supervised** setting (i.e., all ground truth y_t observed during training). The numbers in bold indicate the best results (i.e., our full model of instance attention).

Audiovisual Representation	Classifier			
	MLP	LSTM	AVSDN	AVEL
intra	65.8	66.9	72.3	75.0
intra+inter	66.3	70.4	72.2	75.4
intra+audio	67.2	69.3	73.9	74.8
inter+audio	66.1	66.8	72.1	74.3
Ours	**67.6**	**70.7**	**75.8**	**76.8**

5 Conclusion

We presented a deep learning framework of Audiovisual Transformer, with the ability of cross-modality instance-level attention for audio-visual event localization. Our model jointly exploits intra and inter-frame visual representation while observing audio features, with the self-attention mechanism realized in a transformer-like architecture in supervised or weakly-supervised settings. In addition to promising performances on event localization, our model allows instance-level attention, which is able to attend the proper image region (at the instance level) associated with the sound/event of interest. From our experimental results and ablation studies, the use and design of our proposed framework can be successfully verified.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 109-2634-F-002-037.

References

1. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV. (2018) 1, 4, 5, 6, 8, 9, 10, 11, 12, 13
2. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: ICCV. (2019) 1, 4, 10
3. Ramaswamy, J.: What makes the sound?: A dual-modality interacting network for audio-visual event localization. In: ICASSP. (2020) 1, 4, 10
4. Ramaswamy, J., Das, S.: See the sound, hear the pixels. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). (2020) 1, 4, 10

5. Xuan, H., Zhang, Z., Chen, S., Yang, J., Yan, Y.: Cross-modal attention network for temporal inconsistent audio-visual event localization. In: AAAI. (2020) 1, 4, 10
6. Hu, D., Li, X., et al.: Temporal multimodal learning in audiovisual speech recognition. In: CVPR. (2016) 1
7. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient large-scale multi-modal classification. In: AAAI. (2018) 1
8. Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: CVPR. (2017) 1, 3
9. Wiles, O., Koepke, A.S., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV. (2018) 1, 3
10. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI. (2019) 1, 3
11. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: ICCV. (2019) 1
12. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR. (2016) 1, 3
13. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV. (2018) 2, 3, 4, 6, 7, 8
14. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: CVPR. (2018) 2, 6, 7, 8, 12, 13
15. Arandjelović, R., Zisserman, A.: Objects that sound. In: ECCV. (2018) 2, 3, 6, 7, 8, 12, 13
16. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV. (2018) 2, 3, 6, 7
17. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. TPAMI (2017) 3
18. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.: Large-scale video classification with convolutional neural networks. In: CVPR. (2014) 3
19. Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: ICCV. (2017) 2923–2932 3
20. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. (2016) 3
21. Zolfaghari, M., Singh, K., Brox, T.: ECO: efficient convolutional network for online video understanding. In: ECCV. (2018) 3
22. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV. (2018) 3
23. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV. (2018) 3
24. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: C3D: generic features for video analysis. arXiv (2014) 3, 12
25. Tran, D., Ray, J., Shou, Z., Chang, S., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. CoRR **abs/1708.05038** (2017) 3
26. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR. (2017) 3, 12
27. Lev, G., Sadeh, G., Klein, B., Wolf, L.: RNN fisher vectors for action recognition and image annotation. In: ECCV. (2016) 3

28. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.M.: Videolstm convolves, attends and flows for action recognition. CVIU (2016) 3
29. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: CVPR. (2016) 3034–3042 3
30. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A.: Action recognition with dynamic image networks. TPAMI (2018) 3
31. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. (2017) 3
32. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS. (2016) 3
33. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV. (2016) 3
34. Gao, R., Grauman, K.: 2.5d-visual-sound. CVPR (2019) 3, 4
35. Tian, Y., Guan, C., Goodman, J., Moore, M., Xu, C.: An attempt towards interpretable audio-visual video captioning. arXiv (2018) 3
36. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV. (2018) 3, 4
37. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM TOG (2018) 3
38. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: ICCV. (2019) 4
39. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: ICCV. (2019) 4
40. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV. (2019) 4
41. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: CVPR. (2020) 4
42. Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: ICASSP. (2019) 4, 9, 10
43. Kim, K., Choi, S., Kim, J., Zhang, B.: Multimodal dual attention memory for video story question answering. In: ECCV. (2018) 6
44. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: CVPR. (2018) 6
45. Bai, Y., Fu, J., Zhao, T., Mei, T.: Deep attention neural tensor network for visual question answering. In: ECCV. (2018) 6
46. Shi, Y., Furlanello, T., Zha, S., Anandkumar, A.: Question type guided attention in visual question answering. In: ECCV. (2018) 6
47. Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., Batra, D.: Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In: EMNLP. (2016) 6
48. Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: CVPR. (2017) 6, 8
49. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR. (2018) 6
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. (2017) 6, 8
51. Kim, J., Lee, S., Kwak, D., Heo, M., Kim, J., Ha, J., Zhang, B.: Multimodal residual learning for visual QA. In: NeurIPS. (2016) 8
52. Gemmeke, J.F., Ellis, D.P.W., et al.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. (2017) 9

53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014) 9
54. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. (2009) 9
55. Hershey, S., Chaudhuri, S., et al.: Cnn architectures for large-scale audio classification. In: ICASSP. (2017) 9
56. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997) 12
57. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS. (2014) 12