

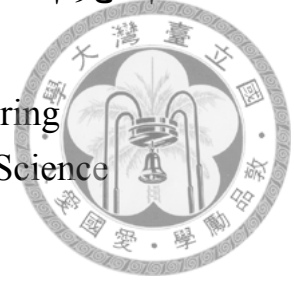
國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering
College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



跨模態共注意視聽事件定位

Cross-Modality Co-Attention for Audio-Visual Event
Localization

林彥伯

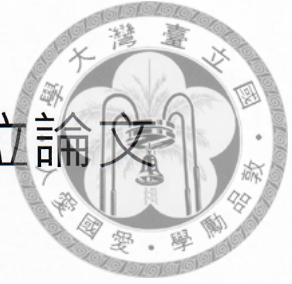
Yan-Bo Lin

指導教授：王鈺強博士

Advisor: Yu-Chiang Frank Wang, Ph.D.

中華民國 108 年 7 月

Jul, 2019



國立臺灣大學 (碩) 博士學位論文

口試委員會審定書

跨模態共注意視聽事件定位

Cross Modality Co-Attention for Audio Visual Event
Localization

本論文係林彥伯君 (R06942048) 在國立臺灣大學電信工程學研究所
完成之碩士學位論文，於民國 108 年 7 月 29 日承下列考試委員審查通
過及口試及格，特此證明

口試委員：

王鈺庚

(簽名)

(指導教授)

邵維辰

林秀宇

所 長

蘇火保

(簽名)



誌謝

感謝指導教授王鈺強老師，在這研究所兩年中的各種指導，老師在研究上讓我有很大彈性自由，有興趣的領域與技術老師都會支持且給予很大的資源，除了學業內容上讓我受益匪淺，許多的建議激發我在思考問題上的深度。在思考一個解決問題的辦法時，還要考慮許多潛在因素和未來所會遇到問題，這是從以前大學時候上課老師一直在強調的事，也是我覺得學生時期最珍貴的收穫，期許在未來的就業跟人生道路上能不忘此道理，在此由衷的感謝老師。也感謝碩士班這兩年課程上遇到的老師，每堂課都讓我學到不同領域的專業，也不吝嗇在課堂外的時間幫忙解惑問題，希望以後有機會我也能充當老師的角色，幫忙有困難的人，並持續得以學習的心態面對未來的挑戰，資訊科技的日新月異，使我不能倦怠。另外也要感謝在求學的路程上陪伴我的很多朋友，包括學弟們、同屆的夥伴、大學的死黨，更重要的是實驗室成員：陳尚甫、陳柏屹、顏嘉緯、李宇哲、劉彥廷、楊福恩、郭冠軒、黃柏翔、李元灝的互相幫忙與學習，還有提供我新穎意見的吼哥賴禹丞，雖然吼哥已經自顧不暇了，還是願意跟我一起做研究，以後再也沒有理由可以煩吼哥，希望吼哥也能趕快完成他的研究順利畢業。感謝這些朋友一起完成了許多事情，還有讓我閒暇之餘不會枯燥乏味，這些都是很珍貴的回憶。更重要的是我的家人，沒有他們，我沒辦法無後顧之憂得順利完成學業，感謝每一個出現在我生活中的人。





Acknowledgements

Thanks to the instructor Professor Yu-Chiang Frank wang for his kind guidance in the two years of the Institute. The teacher has made me have a lot of flexibility in research. The interested areas and technical teachers will support and give great resources. In addition to the academic content, I have benefited a lot, and many suggestions have inspired me to think about the depth of the problem. When thinking about a solution to a problem, there are many potential factors to consider and problems to be encountered in the future. This is something that the teacher has been emphasizing from the previous college class, and it is also the most precious harvest I think during the student period. I hope that in the future employment and life path, I will not forget this truth. I sincerely thank the teacher.

I am also grateful to the teachers I met in the two-year course of the Master's Program. Every class taught me to learn in different fields. I don't want to help solve problems outside the classroom. I hope that I will be able to act as a teacher in the future. Help people with difficulties, and continue to learn the mentality to face the challenges of the future, information technology is changing with each passing day,

Make me not tired. I also want to thank my many friends who accompanied me on my journey to study. Including the younger brothers, the same partners, and the buddies of the university, in addition to the mutual help and study of the laboratory, I have done a lot of things together, and I will not be dull and boring when I am free. These are very precious memories. What's

more important is my family. Without them, I can't finish my studies without any worries. Thanks to everyone who appears in my life.





摘要

視聽事件定位需要人類透過聯合觀察跨模態視聽信息及跨越視頻幀的事件標籤。為了解決這個任務，我們提出了一個跨模式的深度學習框架專注在共同關注視頻事件定位。我們提出的模型能夠利用幀內和幀間時間及視覺信息，以及同時間的音訊資訊，利用觀察上述的三種資訊，來實現共注意視覺物件。搭配視覺，連續觀察到的時間及音訊資訊，我們的模型實現了有新穎的能力來提取空間訊息/時間特徵以改進視聽事件定位。而且，我們的模型能夠產生實例級別的視覺注意力，這將識別圖像最有可能發出聲音的區域/位置，並且在同時有相同物體的場景中找出真正發聲的物體。在實驗設計方面，我們利用了最新穎的方法來跟我們所提出的共注意模組進行比較，並且使用公開的數據集來驗證我們提出方法的有效性，其中我們的實驗結果準確度超過目前現有的方法，可視化的結果也能印證我們提出的架構能達到實例級別的視覺注意力。

關鍵字：視聽特徵, 雙模態, 跨模態, 事件定位, 深度學習, 機器學習, 電腦視覺





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 neural network named Audio-Visual sequence-to-sequence dual network (AVSDN). By jointly taking both audio and visual features at each time segment as inputs, our proposed model learns global and local event information in a sequence to sequence manner. Besides, we also propose a deep learning framework of cross-modality co-attention for audio-visual event localization. The co-attention framework can be applied on existing methods and AVSDN. Our co-attention model 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.

Keywords: Audio-Video Features, Dual Modality, Cross Modality, Event Localization, Deep learning, Machine learning, Computer vision





Contents

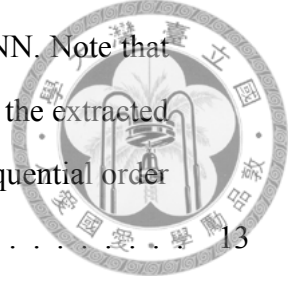
誌謝	iii
Acknowledgements	v
摘要	vii
Abstract	ix
1 Introduction	1
2 Related Work	5
3 Proposed Method	7
3.0.1 Notations and Problem Formulation	7
3.0.2 Audio-Visual Sequence-to-sequence Dual Network (AVSDN) . .	8
3.0.3 Learning Intra and Inter-Frame Visual Representation	12
3.0.4 Cross-Modality Co-Attention for Event Localization	15
4 Experiments	19
4.0.1 Dataset	19
4.0.2 Implementation Details	19
4.0.3 Experiment results	20
4.0.4 Ablation studies	24
5 Conclusion	29





List of Figures

1.1	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.	2
3.1	Overview of our Audio-Visual Sequence-to-sequence Dual Network (AVSDN). Our AVSDN is composed of three main components, which include encoder modules (in orange and blue) for learning visual and audio representations, a fusion network (in purple) for producing global video representation, and a decoder (in green) which jointly takes global and local features for event localization. Note that v and a denote visual and audio features, h and s are the hidden and cell states of LSTMs, and y indicates the event label.	9
3.2	Overview of our proposed Cross-Modality Co-Attention model (CM-CoAtt). Our CM-CoAtt deep learning framework is composed of three main components, intra and inter-frame visual encoders (Sect. 3.2), and a self-attention based mechanism for cross-modality co-attention (Sect. 3.3). Note that \mathbf{v} and \mathbf{a} denote visual and audio features, respectively.	12



3.3 **Intra-frame Visual Encoder:** Local image regions within a video frame are encoded and represented as channels processed by a CNN. Note that a total of R regions is extracted from the input frame, while the extracted regions are fed into a BiLSTM in the forward-backward sequential order for encoding. 13

3.4 **Inter-frame Visual Encoder:** With local image regions described by intra-frame visual encoders, we introduce this inter-frame visual encoder to take three consecutive frames as the inputs, aiming to model short-term temporal information for event localization purposes. Note that $g_{\theta}(\cdot)$ is a simple multilayer perceptron (MLP). 15

3.5 **Cross-modality co-attention:** Observing locally visual-attended features $\tilde{\mathbf{v}}_r^t$ and audio inputs \mathbf{a}^t to output the co-attention features \mathbf{v}_{att}^t across visual, temporal, and audio data domains. 16

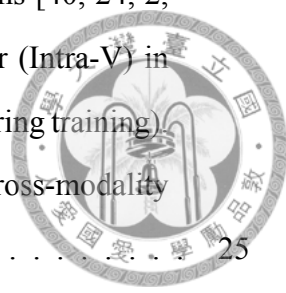
4.1 **Example attention results using AVSDN [28] with our proposed co-attention model:** Each row shows a video input with visually attended regions. Note that the frames bounded in red rectangles denote those with unmatched audio-visual events. It can be seen that our model produced satisfactory attention outputs with the corresponding audio-visual events. 23

4.2 **Example attention results using AVSDN [28] with the attention model of AVEL [40]:** Each row shows a video input with visually attended regions. Note that the frames bounded in red rectangles denote those with unmatched audio-visual events. Take row 3 for example, it can be seen that AVEL would incorrect attend the regions of bell which was not actually associated with the sound of chime. 23



List of Tables

4.1	Comparisons with the state-of-the-art method of [40] in a supervised manner (all ground truth y_t are observed during training. The number in bold indicates the best result.	21
4.2	Comparisons with the state-of-the-art method of [40] in a weakly supervised manner (only ground truth \mathbf{Y} is observed for training). The number in bold indicates the best result.	21
4.3	Performance comparisons using baseline or state-of-the-art localization methods of LSTM, AVEL, and AVSDN and ours in a supervised manner (i.e., all ground truth y_t observed during training). The numbers in bold indicate the best results (i.e., methods with our proposed cross-modality co-attention mechanism).	22
4.4	Performance comparisons using baseline or state-of-the-art localization methods of LSTM, AVEL, and AVSDN and ours in a weakly supervised manner (i.e., only ground truth \mathbf{Y} observed for training). The numbers in bold indicate the best results (i.e., methods with our proposed cross-modality co-attention mechanism).	22
4.5	Ablation studies on our network design, i.e., our decoder taking hidden and cell states of global visual/audio representations as conditioned LSTM inputs. Note that weakly supervised learning is considered in this table.	24



4.6	Comparisons of recent audio-visual co-attention mechanisms [40, 24, 2, 34] with/without integrating our intra-frame visual encoder (Intra-V) 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 cross-modality co-attention).	25
4.7	Comparisons of recent audio-visual co-attention mechanisms [40, 24, 2, 34] with/without our intra-frame visual encoder (Intra-V) in weakly supervised manners (i.e., only ground truth \mathbf{Y} observed during training). The numbers in bold indicate the best results (i.e., with our cross-modality co-attention).	25
4.8	Ablation studies on the exploiting inter-frame visual information (Inter-V) in different temporal relational modules. Note that fully supervised settings are considered in this table, and the nubmers in bold indicate the best performances.	27



Chapter 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 [15, 34, 49, 1, 39, 40] attracts the attention from computer vision and machine learning communities. As depicted in Fig. 1.1, 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 [21, 47, 23, 15]. However, existing models require the presence of both visual and audio information to learn the event of interest. In other words, they cannot deal with scenarios with partial information observed or in weakly supervised settings.

On the other hand, cross-modal synthesis models [52, 8] have also been proposed to exploit information observed from different data modalities. For example, [45, 51] are capable of converting spoken audio data with captions into face video frames. However, since these models require label annotation for each data domain, and the synthesized videos are typically limited (e.g., for cropped-out face regions), such techniques cannot

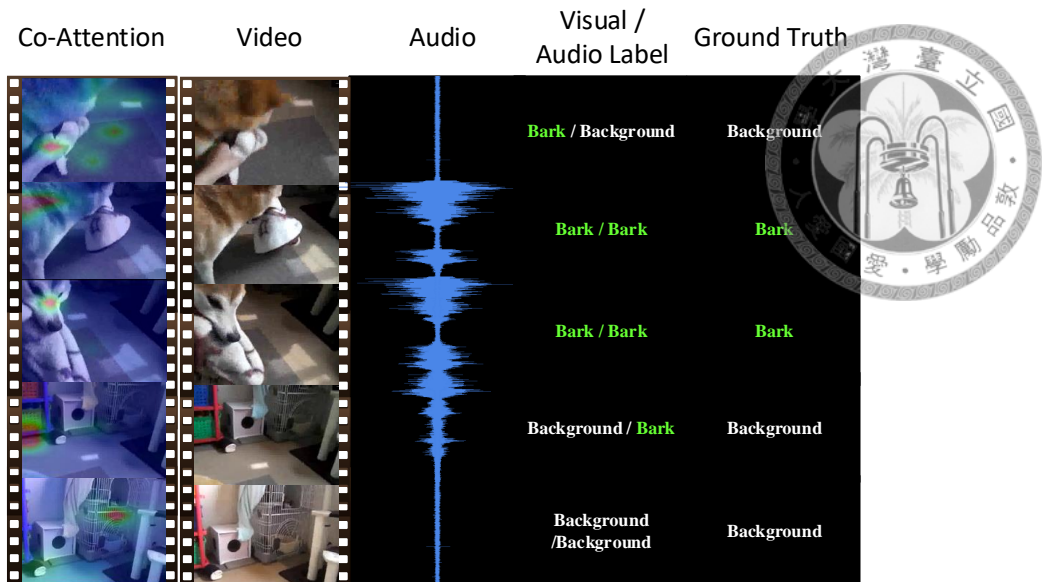


Figure 1.1: Illustration of audio-visual event localization (recognizing video event with matched visual and audio information). Note that the first column shows our correct localization outputs with cross-modality co-attention, the 2nd and 3rd columns show the video and audio inputs across five consecutive frames, with ground truth visual/audio and event labels depicted in the last two columns.

be easily extended to audio-visual event localization for general video data.

To address this challenging task, we propose an end-to-end deep learning framework of Audio-Visual sequence-to-sequence dual network (AVSDN). Based on sequence to sequence (seq2seq) [38] and autoencoder, our network architecture takes both audio and visual data at each time segment as inputs and exploits global and local event information in a seq2seq manner. More importantly, our model can be learned in a fully or weakly supervised settings, i.e., ground truth event labels observed in the frame or video levels.

Moreover, to better associate audio and visual information for video event localization, we propose a novel deep attention model which jointly performs visual, temporal, and audio cross-modality co-attention for video event localization. This is realized by advancing LSTMs for encoding intra-frame patches, followed by exploitation of encoded intra-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. 1.1). 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. We now highlight the contributions of this work as follows:

- We propose a unique end-to-end trainable network for audio-visual event localization, which is able to jointly take visual and audio data as inputs. By exploiting this cross-modality information across time, such encoded global features will be conditioned on the decoder for event localization.
- We propose a novel end-to-end trainable module for visual, temporal, and audio co-attention, which can exploit cross-modality information across video frames for event localization.
- Without attention supervision, our model performs instance-level attention during event localization. This is achieved by advancing LSTMs to model local image regions within each frame, followed by temporal and audio information jointly exploited across video frames.
- Experimental and visualization results demonstrate that our proposed module performs favorably against state-of-the-art approaches in both fully and weakly supervised settings.





Chapter 2

Related Work

Video Classification. Methods based on deep neural networks have shown promising performances on the task of video classification [12, 22, 53, 44, 54, 46, 50], 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 [41, 42, 7]. On the other hand, long short term memory (LSTM) networks [12] has also been employed to observe 2D CNN features over time. Such recurrent neural networks (RNN) [12, 26, 27] 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 [26, 27], some works choose to sample frames from the entire video to learn robust reasoning relational representation [50, 44, 54, 6, 5].

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 [1, 3, 32, 31, 15, 39]. For example, Aytar et al. [3] learn the joint representation from audio-visual data, with the goal to identifying the content using data in either modality. Arandjelovic and Zisserman [1] also exploit the variety of audio-visual information for learning better representation in audio-visual

correspondence tasks. Furthermore, they [2] visualize sound localization in visual scenes, which would serve as the bridge connecting between audio and visual modality. Owens et al. [30] 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 the locations of sound sources, and on/off-screen source separation. These studies [30, 2] 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 [9, 45, 51] 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 [30, 49, 14, 15, 13] also utilize visual representation to solve audio source separation and denoising. 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). 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.



Chapter 3

Proposed Method

3.0.1 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.

For the sake of completeness, we first define the settings and notations which will be used in this paper. Following [40], 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-wise labels are available as $\mathbf{y}^t = \{y_k^t | y_k^t \in \{0, 1\}, \sum_{k=1}^{C+1} y_k^t = 1, t \in \mathbb{N}\}$, $\mathbf{y}^t \in \mathbb{R}^{C+1}$, 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 have access to the video-level event labels during the training phase. Note that the video-level event labels are

processed by averaging the segment event labels $\mathbf{Y} = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t$, $\mathbf{Y} \in \mathbb{R}^{C+1}$. 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.

Our proposed framework consists two major parts: AVSDN [28] and CMCoAtt. As shown in Fig 3.1, our AVSDN [28] is composed of three components: encoder modules for learning visual and audio representations, fusion network for producing global video representation, and decoder to take global and local features for event localization.

As shown in Fig 3.2, our CMCoAtt consists of network modules of intra and inter-frame visual encoder. Together with audio features across video frames, cross-modality co-attention can be performed. That is, our model allows joint attention over visual, temporal and audio domains for audio-visual event localization. We now discuss these two modules in the following subsections.

3.0.2 Audio-Visual Sequence-to-sequence Dual Network (AVSDN)

To address the audio-visual event localization problem in both supervised and weakly-supervised problems, we propose a novel framework named Audio-Visual sequence-to-sequence dual network (AVSDN). Based on seq2seq [38] and autoencoder, our network architecture takes both audio and visual data at each time segment as inputs and exploits global and local event information in a seq2seq manner.

As shown in Fig 3.1, the framework is composed of three components: encoder modules for learning visual and audio representations, fusion network for producing global video representation, and decoder to take global and local features for event localization. More details about the three components in AVSDN can be obtained as follows:

Encoder: learning global visual and audio representations. The encoder, which is the first part of our network AVSDN, is aimed to extract global visual and audio representations for fusion network.

Before learning the global features, we have to obtain the segment visual and audio representations by utilizing CNNs. To better learn the visual and audio embedding features,

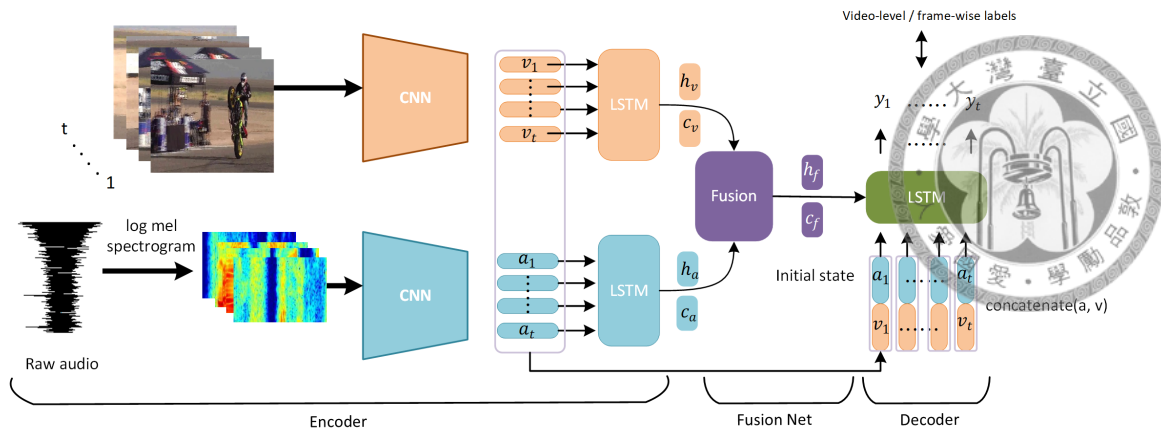


Figure 3.1: **Overview of our Audio-Visual Sequence-to-sequence Dual Network (AVSDN).** Our AVSDN is composed of three main components, which include encoder modules (in orange and blue) for learning visual and audio representations, a fusion network (in purple) for producing global video representation, and a decoder (in green) which jointly takes global and local features for event localization. Note that v and a denote visual and audio features, h and s are the hidden and cell states of LSTMs, and y indicates the event label.

our CNNs are learned from the large-scale dataset (ImageNet [11] and AudioSet [16]) which are highly shown useful for vision and audition tasks. To be specific, for visual frames we sample a frame and obtain the visual representation from pre-trained ResNet-152 [18], which has been trained on ImageNet. On the other hand, for one raw audio, we convert one segment to log mel spectrogram and extract an audio representation each 1s from VGGish [19] trained on AudioSet.

In order to further learn the global visual and audio representations, we now utilize the Long Short-Term Memory (LSTM) [20], which is known to exploit long-range temporal dependencies, to generate encoded temporal representation sequence. Generally, the most common implementation of vanilla LSTM [17] includes various gate mechanisms such as input gate, forget gate, output gate, memory state, and hidden state etc. The utilized

LSTM unit in our proposed model is illustrated in Eq. (3.1).

$$\begin{aligned}
 f_t &= \sigma_g(W_f x_t + U_f x_t + b_f) \\
 i_t &= \sigma_g(W_i x_t + U_i x_t + b_i) \\
 o_t &= \sigma_g(W_o x_t + U_o x_t + b_o) \\
 c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
 h_t &= o_t \circ \sigma_h(c_t)
 \end{aligned} \tag{3.1}$$



Each time step in Eq. (3.1) denotes subscript t , and $x_t \in \mathbb{R}^d$ denotes the given input of time step t . $f_t \in \mathbb{R}^h$, $i_t \in \mathbb{R}^h$ and $o_t \in \mathbb{R}^h$ are forget, input and output gate's activation vector respectively. $h_t \in \mathbb{R}^h$ and $c_t \in \mathbb{R}^h$ are hidden and cell state of the LSTM unit. When $t = 0$, $c_0 = 0$ and $h_0 = 0$ would be the initial values. $W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$ are weight matrices and bias vector which can be learned during train phase. Where d and h refer the number of input features and number of hidden units. Element-wise product is denoted by \circ . Activation function: σ_g is sigmoid function and σ_c is hyperbolic tangent function.

Eventually, all the audio and visual segments are the inputs of the two designed LSTM (audio and video separately). Hence, the last time step T of hidden and cell state can be generated as the global representations of audio and visual tracks.

Fusion network: learning video event representation.

After obtaining the audio and visual global representations, our goal is to convert these two representations into one video event representation. To perform such a fusion mechanism, our fusion network is designed to fuse cross-modality features which are built based on dual multimodal residual network (DMRN) fusion block [40]. As mentioned above, the last time step T of hidden and cell state from encoder can be given as the representations of audio and visual tracks. Following [40], with time step T , audio and visual hidden state (h_T^a, h_T^v) and cell state (c_T^a, c_T^v) can be fused with Eq.(3.2). After fusion hidden and cell state, these fused state will be the initial state of the decoder LSTM (one LSTM of

right-half Fig.3.2).

$$\begin{aligned}
 h_T^{a'} &= \sigma_c(h_T^a + \frac{1}{2}(g_\theta(h_T^a) + g_\theta(h_T^v))) \\
 h_T^{v'} &= \sigma_c(h_T^v + \frac{1}{2}(g_\theta(h_T^a) + g_\theta(h_T^v))) \\
 h_T^f &= h_T^{a'} + h_T^{v'}
 \end{aligned} \tag{3.2}$$



where σ_c is denoted as hyperbolic tangent function and $g_\theta(\cdot)$ is multilayer perceptrons (MLP) with parameters θ . The cell states can be fused like hidden states. The fused states, h_t^f and c_t^f , turn to be the initial states in our decoder. Because compared with vanilla LSTM, a representative initial state can benefit a LSTM for prediction [38]. Thus, we take the fused states for the initialization for the decoder.

Decoder: localization of video events using global and local cross-modality representations. Generally, our decoder is aimed to perform the supervised and weakly-supervised event localization. Thus, given both the fused global representations from the fused network and local features of audio and video, our decoder will generate the corresponding labels segment-wisely. The architecture of our decoder is a single LSTM. Different from each encoder, the inputs of the decoder are concatenated features which are global and local cross-modality representations. We concatenate a_t and v_t which are audio and visual segment features from pre-trained CNN. Our decoder is designed to not only learn spatial cross-modality representations but temporal ones. Especially weakly supervised setting, we can only access to the video-level labels in the training phase. All the individual predictions will be aggregated by average pooling in Eq.(3.3),

$$\hat{m} = avg(m_1, m_2, \dots, m_T) = \frac{1}{T} \sum_{t=1}^T m_t, \tag{3.3}$$

where m_1, \dots, m_T are the predictions from the last fully connected layer of our model. The average prediction \hat{m} over softmax function can be the probability distribution of the event category. For both the weakly-supervised and supervised setting, the predicted probability distribution can be optimized by video-level labels through binary cross-entropy.

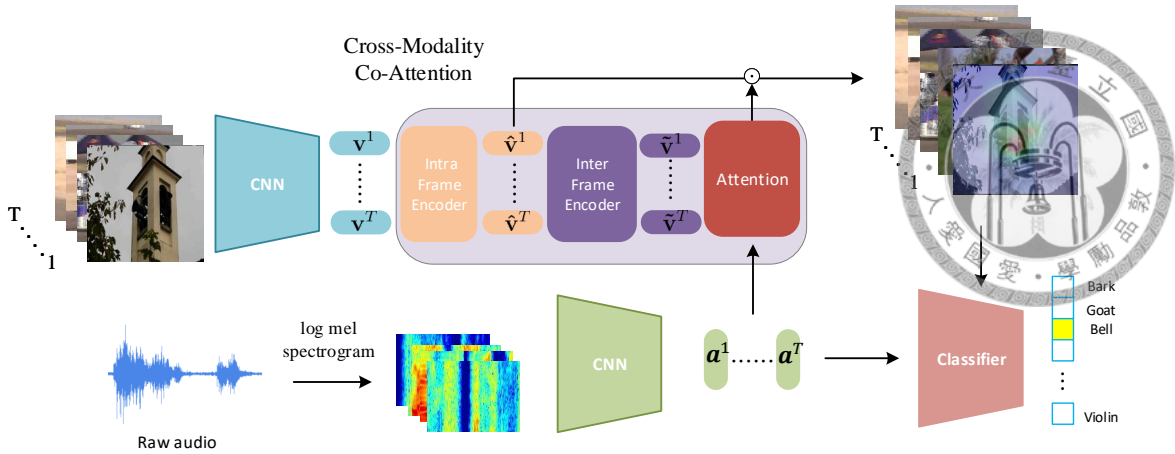


Figure 3.2: **Overview of our proposed Cross-Modality Co-Attention model (CM-CoAtt).** Our CM-CoAtt deep learning framework is composed of three main components, intra and inter-frame visual encoders (Sect. 3.2), and a self-attention based mechanism for cross-modality co-attention (Sect. 3.3). Note that \mathbf{v} and \mathbf{a} denote visual and audio features, respectively.

3.0.3 Learning Intra and Inter-Frame Visual Representation

Intra-frame visual encoder . Visual attention has been widely utilized in recent VQA and audio-visual related tasks [25, 29, 4, 35, 10, 40, 34, 2]. 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., supervision or guidance at the network outputs) [34, 2, 49, 30, 10].

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.

Inspired by [48], we utilize recurrent neural networks 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 a video frame into a bidirectional LSTM [20] network, which encode the image patches of that frame in a sequential yet bidirectional order. To be more specific, as illustrated in Fig 3.3, we divide a input video frame at time step t into R patches, and extract the CNN feature for

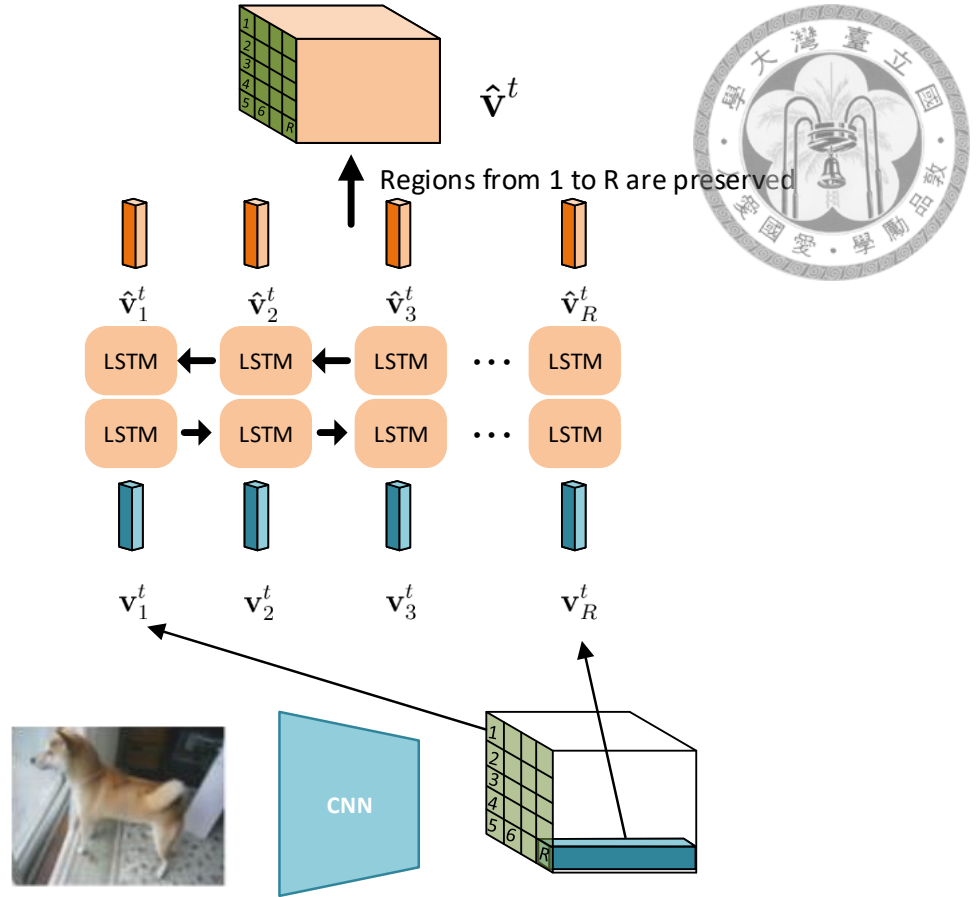


Figure 3.3: **Intra-frame Visual Encoder**: Local image regions within a video frame are encoded and represented as channels processed by a CNN. Note that a total of R regions is extracted from the input frame, while the extracted regions are fed into a BiLSTM in the forward-backward sequential order for encoding.

each patch. These visual representations of each region are denoted as $\{\mathbf{v}_r^t, r = 1, 2, \dots, R\}$, where $\mathbf{v}_r^t \in \mathbb{R}^{1 \times K}$ represents the visual features of the r th patch. These visual features are served as the inputs to the bidirectional LSTMs, which is described below:

$$\hat{\mathbf{v}}_r^t = LSTM^f(\mathbf{v}_r^t) + LSTM^b(\mathbf{v}_r^t), \quad (3.4)$$

where $LSTM^f(\cdot)$ and $LSTM^b(\cdot)$ denote the forward and backward LSTMs, respectively, and $\hat{\mathbf{v}}_r^t$ indicates intra-frame visual representations for r th patch. We gather R patches for intra-frame representations of video frame at time t , that is, $\hat{\mathbf{V}}^t = \{\hat{\mathbf{v}}_1^t, \dots, \hat{\mathbf{v}}_R^t\} \in \mathbb{R}^{R \times K}$.

It can be seen that, via advancing recurrent neural networks, visual representation encoded in this stage describes local spatial information within a video frame. By combing

temporal and audio information in the following stages, this intra-frame visual encoder allows improved attention at the instance-level as later verified.



Inter-frame visual encoder . With image representation preserving intra-frame local visual information obtained, inter-frame visual information needs to be exploited so that relational reasoning across video frames can be performed. In a recent work of [33], such strategies have shown promising performances on VQA tasks. Similarly, the network module of temporal relational reasoning was presented in [50], which captures temporal relational information for activity recognition.

To encode visual information across video frames for event localization, we need to integrate temporal information into the derived $\hat{\mathbf{v}}^t$. In other words, we need to exploit inter-frame visual information based on the intra-frame ones previously produced. In our proposed framework, we choose to perform this inter-frame encoding by feeding consecutive video frames (in terms of $\hat{\mathbf{v}}^t$) into an encoder. As depicted in Fig 3.4, the encoded inter-frame visual representation is derived as:

$$\tilde{\mathbf{v}}^t = g_\theta(\hat{\mathbf{v}}^{t-1}, \hat{\mathbf{v}}^t, \hat{\mathbf{v}}^{t+1}), \quad (3.5)$$

where $\hat{\mathbf{v}}^t \in \mathbb{R}^{R \times K}$ is the encoded intra-frame feature representation at time step t derived by (3.4). Note that $g_\theta(\cdot)$ is a standard multilayer perceptron with parameters θ to perform inter-frame visual encoding. Thus, the output $\tilde{\mathbf{v}}^t \in \mathbb{R}^{R \times D}$ not only contains the regional relationship within a image but also temporal differences is jointly explored.

A final remark on our inter-frame visual encoder is that, the use of simple MLP-based encoders allows us to exploit short-term temporal information across locally attended visual features across frames. Together with the audio inputs, it would be sufficient to attend and recognize audio-visual events in a video. We do *not* consider LSTM-based modules to exploit long-term temporal information, which would obviously increase model complexity and training difficulty.

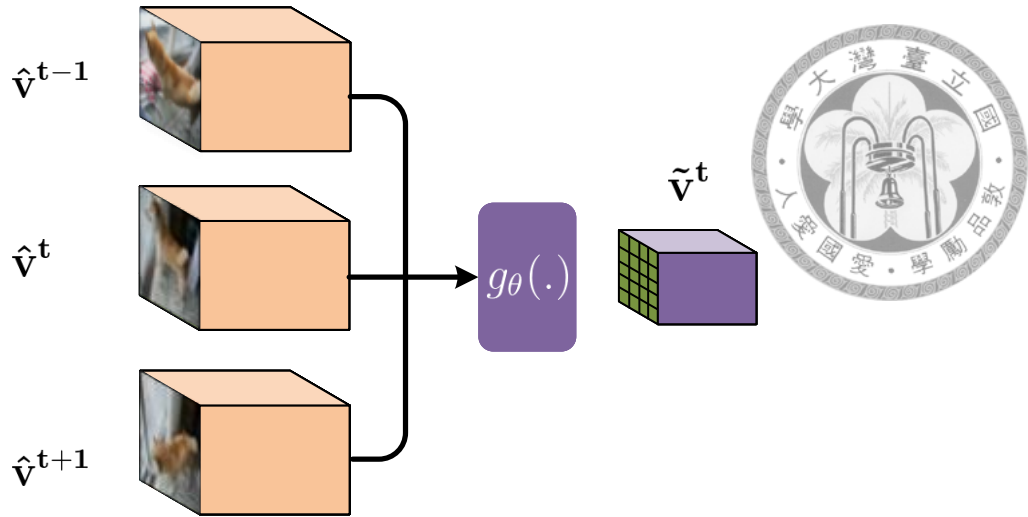


Figure 3.4: **Inter-frame Visual Encoder**: With local image regions described by intra-frame visual encoders, we introduce this inter-frame visual encoder to take three consecutive frames as the inputs, aiming to model short-term temporal information for event localization purposes. Note that $g_{\theta}(\cdot)$ is a simple multilayer perceptron (MLP).

3.0.4 Cross-Modality Co-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 [34, 2, 49, 30] 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 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.

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,

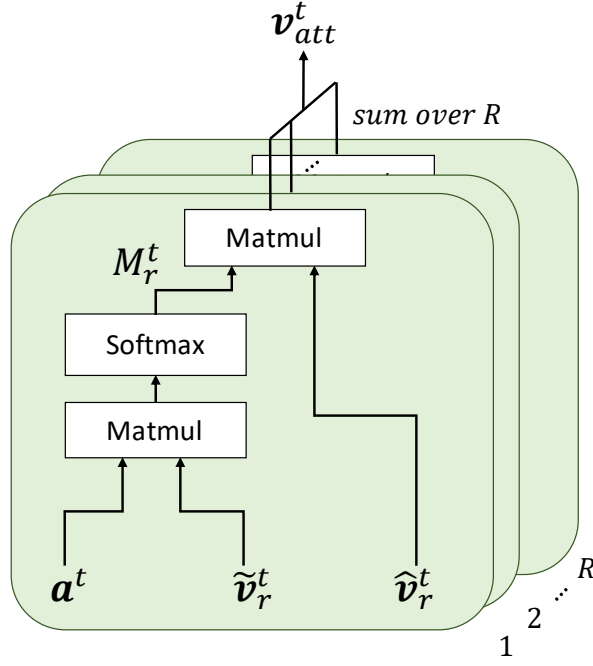


Figure 3.5: **Cross-modality co-attention**: Observing locally visual-attended features $\hat{\mathbf{v}}_r^t$ and audio inputs \mathbf{a}^t to output the co-attention features \mathbf{v}_{att}^t across visual, temporal, and audio data domains.

which would make the localization of audio-visual events more applicable. To achieve this goal, we advance the concept of self-attention [43] for computing a soft confidence score map, indicating the correlation between the attended visual and audio features. Different from existing co-attention mechanisms like [40, 34, 2, 49, 48, 24], 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. 3.5, we have the r th local visual feature of time t $\{\tilde{\mathbf{v}}_r^t, r = 1, 2, \dots, R\}$, and our co-attention model aims to produce the weight to depict how relevant $\hat{\mathbf{v}}_r^t$ 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:

$$M_r^t = \text{Softmax}(\tilde{\mathbf{v}}_r^t \cdot (\mathbf{a}^t)'), \quad (3.6)$$

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.,

$$\mathbf{v}_{att}^t = \sum_{r=1}^R M_r^t \hat{\mathbf{v}}_r^t. \quad (3.7)$$

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., [40, 28]). We will detail this implementation and provide thorough comparisons in the experiment section.





Chapter 4

Experiments

4.0.1 Dataset

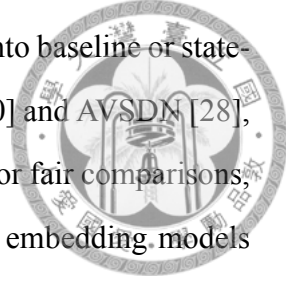
For the audio-visual event localization, we follow [40] and consider the *Audio-Visual Event* (AVE) [40] dataset (a subset of Audioset [16]) for experiments (e.g., Church bell, Dog barking, Truck, Bus, Clock, Violin, etc.). This AVE dataset includes 4143 videos with 28 categories, and audio-visual labels are annotated at every second.

4.0.2 Implementation Details

In this section, we present the implement details about the evaluation frameworks. For visual embedding, we utilize the ResNet151 [18] pre-trained on ImageNet [11] to extract 2048-dimensional visual feature for each frame. The feature map of whole video frames with T seconds long is $\mathbb{R}^{T \times 7 \times 7 \times 2048}$. We obtain 7×7 channels and 2048 dimension with each channel. Each channel is processed by multilayer perceptrons (MLP) into 512 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 [19], which is pre-trained on AudioSet [16]. 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 per-

centage of correct matchings over all test input frames as the prediction accuracy. In this paper, we feed our cross-modality co-attention visual representation into baseline or state-of-the-art audio-visual classifiers, including a naive LSTM, AVEL [40] and AVSDN [28], to verify the effectiveness of our co-attended features. We note that, for fair comparisons, we apply ResNet-151 as the visual backbone and VGGish as audio embedding models when considering AVEL [40] and AVSDN [28] in our experiments.



4.0.3 Experiment results

Audio-Visual sequence-to-sequencedual network (AVSDN). In AVSDN [28], we use different visual pre-trained embedding compared with Tian et al [40]. Visual pre-trained embedding in Tian et al is VGG16 [37]. Thus, we re-implement the model with ResNet-151 [18] visual pre-trained embedding and show each one modality results. In a fully supervised manner, all the frame-wise labels are used during training. In Table 4.1, our model has better results compared with state-of-the-art methods even if the model [40] is with cross-modality attention mechanism [43] which can find the audio location in the video scene [40, 2]; In a weakly supervised manner, Table 4.2 shows our model outperforms other methods as well.

Furthermore, we further exploit our **Cross-Modality CoAttention model (CMCoAtt)** with different classifier. In Table 4.3, we compare the performance of supervised event localization using baseline and recent models with and without our cross-modality co-attention features. As for the weakly supervised setting, we repeat the same experiments and list the performance comparisons in Table 4.4. From both tables presented, it is clear that use of our cross-modality co-attended features would increase the localization accuracy. In other words, either observing frame-level or video-level labels, our proposed co-attention model would properly extract cross-modality features for improved audio-visual event localization.

We now present example visualization results in fully supervised settings using the AVSDN classifier. The attention output produced by ours and the method of [40] are shown in Fig. 4.1 and 4.2, respectively. We note that, the first and second rows in these two figures showed scenes with multiple objects, but only one or few of the objects were

Table 4.1: Comparisons with the state-of-the-art method of [40] in a **supervised** manner (all ground truth y_t are observed during training. The number in bold indicates the best result.

Method	Accuracy (%)	Remarks
AVEL [40]	59.5	audio only
	55.3	visual only (VGG16)
	71.4	audio+visual (VGG16)
	72.7	audio+visual w/ att (VGG16)
	65.0	visual only (ResNet-151)
	74.0	audio+visual (ResNet-151)
	74.7	audio+visual w/ att (ResNet-151)
AVSDN [28]	75.4	audio+visual (ResNet-151)



Table 4.2: Comparisons with the state-of-the-art method of [40] in a **weakly supervised** manner (only ground truth \mathbf{Y} is observed for training). The number in bold indicates the best result.

Method	Accuracy (%)	Remarks
AVEL[40]	53.4	audio only
	52.9	visual only (VGG16)
	63.7	audio+visual (VGG16)
	66.7	audio+visual w/ att (VGG16)
	63.4	visual only (ResNet-151)
	71.6	audio+visual (ResNet-151)
	73.3	audio+visual w/ att (ResNet-151)
AVSDN [28]	74.2	audio+visual (ResNet-151)

associated with the sound of interest. In the first row, airplanes and cars were potential objects which would result in engine sounds. Since only the airplane was moving across video frames, such characteristic was successfully captured by our cross-modality co-attention model. Similar remarks can be applied to the results in the second rows of these two figures, in which only one of the persons was playing violin. Without our intra and inter-frame visual information, direct association of visual and audio features would not provide satisfactory attention outputs (as shown in Fig. 4.2).

As for the third row in these two figures, the bell was chiming in the last two frames.

Table 4.3: Performance comparisons using baseline or state-of-the-art localization methods of LSTM, AVEL, and AVSDN and ours in a **supervised** manner (i.e., all ground truth y_t observed during training). The numbers in bold indicate the best results (i.e., methods with our proposed cross-modality co-attention mechanism).

Method	Accuracy (%)	CM-CoAtt
LSTM	74.98 75.82	N Y
AVEL[40]	74.00 76.37	N Y
AVSDN [28]	75.4 77.86	N Y

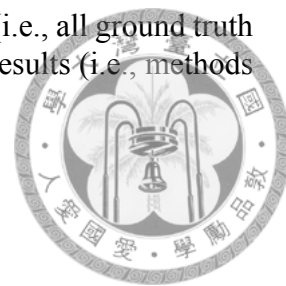


Table 4.4: Performance comparisons using baseline or state-of-the-art localization methods of LSTM, AVEL, and AVSDN and ours in a **weakly supervised** manner (i.e., only ground truth \mathbf{Y} observed for training). The numbers in bold indicate the best results (i.e., methods with our proposed cross-modality co-attention mechanism).

Method	Accuracy (%)	CM-CoAtt
LSTM	73.11 73.81	N Y
AVEL[40]	71.60 74.30	N Y
AVSDN [28]	74.20 75.85	N Y

It can be seen that our model did not attend on the bell region and identify the event until these two frames. It again verifies that our cross-modality co-attention is capable of discerning relation between the audio signal and visual image regions. As for the last two rows, the supervision of inter-frame allows us to concisely identify the region of interest, comparing to the direct use of [40] for attention.

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.

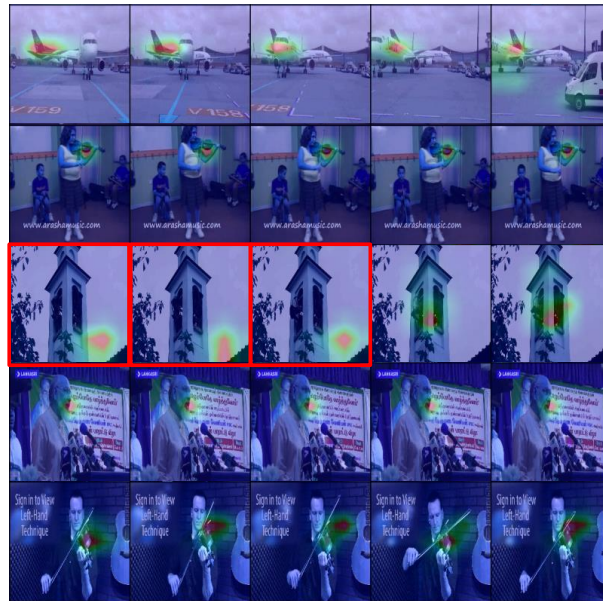


Figure 4.1: **Example attention results using AVSDN [28] with our proposed co-attention model:** Each row shows a video input with visually attended regions. Note that the frames bounded in red rectangles denote those with unmatched audio-visual events. It can be seen that our model produced satisfactory attention outputs with the corresponding audio-visual events.

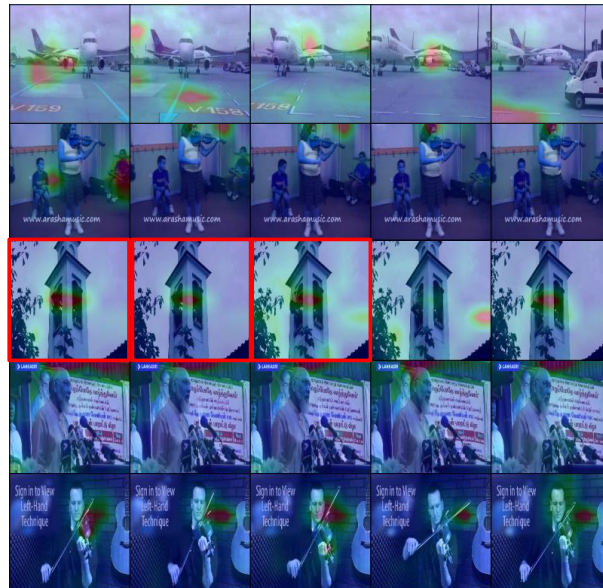
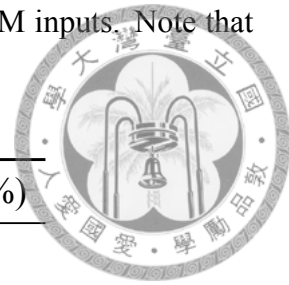


Figure 4.2: **Example attention results using AVSDN [28] with the attention model of AVEL [40]:** Each row shows a video input with visually attended regions. Note that the frames bounded in red rectangles denote those with unmatched audio-visual events. Take row 3 for example, it can be seen that AVEL would incorrect attend the regions of bell which was not actually associated with the sound of chime.

Table 4.5: Ablation studies on our network design, i.e., our decoder taking hidden and cell states of global visual/audio representations as conditioned LSTM inputs. Note that weakly supervised learning is considered in this table.

Method	Accuracy (%)
Visual input only	70.2
Audio input only	70.9
Audio + Visual (label guided)	72.6
Fusion (Ours)	74.2



4.0.4 Ablation studies

In this section, we will verify the design and contributions of different global information for the decoder LSTM in AVSDN [28]. Besides, we also verify our intra and inter-frame visual encoders by using co-attention mechanisms [40, 24, 2, 34]. This would support the learning and exploitation of intra or inter-frame visual representation for audio-visual event localization.

Global representation in AVSDN. In Table 4.5, first-two rows show the results which initial state is only from visual or audio content respectively. Further, we want to explore whether encoder LSTM can learn global event information or not. With the last hidden states from visual and audio modality individually, these hidden states are guided by video-level labels through a simple multilayer perceptron (MLP). The third row in Table 4.5 shows the result is improved compared with only one modality. However, extra loss functions for guiding the last hidden states are not needed. The last hidden states are well-learned during the training of our AVSDN [28].

Intra-frame visual representation. We note that, existing co-attention mechanisms typically operate at each single frame while associating visual and audio information. Recall that we have notations that $\mathbf{v}^1 \in \mathbb{R}^{49 \times 128}$ denotes visual feature, $\mathbf{v}_r \in \mathbb{R}^{1 \times 128}$ denotes

¹note that we omit the superscript t which indicates time step for simplicity here

Table 4.6: Comparisons of recent audio-visual co-attention mechanisms [40, 24, 2, 34] with/without integrating our intra-frame visual encoder (Intra-V) 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 cross-modality co-attention).

Attention Mechanism	Method		Intra-V
	AVEL [40]	AVSDN [28]	
None	74.00	75.40	N
Add [40, 24]	74.70	76.12	N
	75.04	75.57	Y
Dot [2, 34]	71.06	75.21	N
	75.54	75.57	Y
CM-CoAtt (Ours)	75.55	76.02	N
	76.37	77.86	Y

Table 4.7: Comparisons of recent audio-visual co-attention mechanisms [40, 24, 2, 34] with/without our intra-frame visual encoder (Intra-V) in **weakly supervised** manners (i.e., only ground truth \mathbf{Y} observed during training). The numbers in bold indicate the best results (i.e., with our cross-modality co-attention).

Attention Mechanism	Method		Intra-V
	AVEL [40]	AVSDN [28]	
None	71.60	74.20	N
Add [40, 24]	73.30	74.68	N
	74.08	73.81	Y
Dot [2, 34]	72.64	72.29	N
	73.75	72.64	Y
CM-CoAtt (Ours)	73.66	75.15	N
	74.30	75.85	Y

visual feature of the r th region and $\mathbf{a} \in \mathbb{R}^{1 \times 128}$ denotes audio feature. One way to compute the attention map is to directly measure the attention score through inner products between \mathbf{v}_r and \mathbf{a} for every r , which produces an attention map $M \in \mathbb{R}^{1 \times 49}$. This type of attention can be interpreted as calculating the cosine similarity between the visual and audio features for each video frame, while taking their similarities as the attention weights.

The other way to generate attention map is to add \mathbf{a} on each \mathbf{v}_r , and feed the added features as the inputs to a MLP. Then, this MLP would output an attention map $M \in$

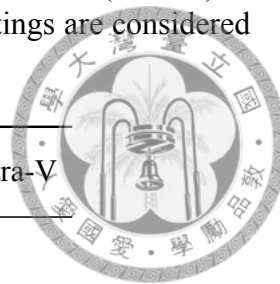
$\mathbb{R}^{1 \times 49}$. Both of these two co-attention methods can jointly work with our intra-frame visual encoding component as long as \mathbf{v} is replaced with intra-frame visual feature. In Table 4.6 and 4.7, we show and compare these two types of co-attention mechanisms (denoted as dot and add respectively) with our full model, with and without utilizing the intra-frame visual features (denoted as Intra-V). More specifically, we apply three types of co-attention (Dot, Add, CM-CoAtt) along with the baseline model without any co-attention mechanism. For each method of interest, two types of classifiers (AVEL, AVSDN) are deployed.

From the final results listed in Table 4.6 (for supervised settings) and Table 4.7 (for weakly supervised settings), we can see that the exploitation of intra-frame visual features for encoding and observing local image regions would be preferable. The main reason is that dot-based co-attention directly computes the attention scores between visual regions and audio feature. It implies that the attention scores are simply determined based on the semantic relation between audio feature and image features. Our intra-frame visual representation additionally exploits local image regions for observing local and consecutive semantic information, and thus it would contain more information during the attention process. We note that, intra-frame visual features using AVSDN resulted in slightly degraded performances. This is probably due to the fact that it is generally more difficult to train additional network models with more parameters like LSTMs for calculating attention scores. Nevertheless, from the above results, we can confirm that the exploitation of intra-frame visual features for encoding and observing local image regions would be preferable.

Inter-frame visual representation. As to study the effects of learning inter-frame visual representations for cross-modality co-attention, we consider different methods to model such inter-frame visual features. To model across frames visual representation, we utilize 3D convolutional networks [41] (Conv3D) and LSTM [20] network in our work. We note that, for standard Convolutional Neural Network [36] and the recent I3D Network [7], 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

Table 4.8: Ablation studies on the exploiting inter-frame visual information (Inter-V) in different temporal relational modules. Note that fully supervised settings are considered in this table, and the numbrs in bold indicate the best performances.

Temporal mechanism	Method		Intra-V
	AVEL [40]	AVSDN [28]	
Conv3D-add	74.88	75.87	N
	74.98	76.14	Y
Conv3D-dot	74.50	75.50	N
	75.70	75.40	Y
LSTM-add	74.50	75.12	N
	75.07	76.62	Y
LSTM-dot	73.08	74.90	N
	75.15	76.24	Y
CM-CoAtt (Ours)	75.55	76.02	N
	76.37	77.86	Y



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 Conv3D and LSTM are able to be utilized in current co-attention [40, 24, 34, 2] methods. There are two typical co-attention mechanisms: add and dot co-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 4.8, our cross-modality co-attention performs favorably against other models with inter-frame visual encoding. In this table, the suffix of temporal mechanism is the co-attention method (e.g., add [40, 24] and dot [34, 2]). 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 MLPs, whose computation cost is lower than the models using Conv3D and LSTM. Based on the above results and observations, we can also confirm the learning of inter-frame visual features would be preferable in our cross-modality co-attention model, which would result in satisfactory event localization performances.





Chapter 5

Conclusion

In this work, we present Audio-Visual sequence-to-sequence dual network (AVSDN) for video event localization, which can be learned in fully or weakly supervised fashions. Our network takes both audio and visual local features, together with integrated global representation, to perform event localization in a sequence to sequence manner.

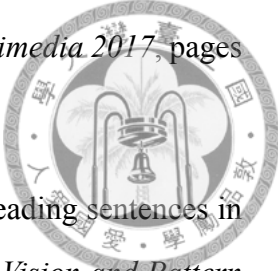
Besides, we presented a deep learning framework for cross-modality co-attention which can be applied on current method and our AVSDN [28], with the goal of addressing the task of audio-visual event localization in fully or weakly supervised learning settings. Our model jointly exploits intra and inter-frame visual representation while observing audio features. Together with a self-attention based mechanism, co-attention across the above feature modalities can be performed. In addition to promising performances on event localization, our model additionally 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.





Bibliography

- [1] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [2] R. Arandjelović and A. Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [3] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [4] Y. Bai, J. Fu, T. Zhao, and T. Mei. Deep attention neural tensor network for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [5] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.
- [6] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3034–3042, 2016.
- [7] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

- 
- [8] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017.
- [9] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [10] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [12] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.
- [13] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 2018.
- [14] 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)*, 2018.
- [15] R. Gao and K. Grauman. 2.5d-visual-sound. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

- [16] J. F. Gemmeke, D. P. W. Ellis, et al. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [17] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *CoRR*, abs/1503.04069, 2015.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] S. Hershey, S. Chaudhuri, et al. Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [20] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9:1735–80, 12 1997.
- [21] D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [23] D. Kiela, E. Grave, A. Joulin, and T. Mikolov. Efficient large-scale multi-modal classification. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [24] J. Kim, S. Lee, D. Kwak, M. Heo, J. Kim, J. Ha, and B. Zhang. Multimodal residual learning for visual QA. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.



- [25] K. Kim, S. Choi, J. Kim, and B. Zhang. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [26] G. Lev, G. Sadeh, B. Klein, and L. Wolf. RNN fisher vectors for action recognition and image annotation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 833–850, 2016.
- [27] Z. Li, K. Gavriluyk, E. Gavves, M. Jain, and C. G. M. Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding (CVIU)*.
- [28] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [29] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [30] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multi-sensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [31] 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)*, 2016.
- [32] 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)*, 2016.

- [33] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [34] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [35] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [36] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576. Curran Associates, Inc., 2014.
- [37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [38] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [39] Y. Tian, C. Guan, J. Goodman, M. Moore, and C. Xu. An attempt towards interpretable audio-visual video captioning. *CoRR*, abs/1812.02872, 2018.
- [40] 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)*, 2018.
- [41] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.
- [42] D. Tran, J. Ray, Z. Shou, S. Chang, and M. Paluri. Convnet architecture search for spatiotemporal feature learning. *CoRR*, abs/1708.05038, 2017.

- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [44] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [45] O. Wiles, A. S. Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [46] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [47] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [48] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [49] 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)*, 2018.
- [50] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [51] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

- [52] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [53] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2923–2932, 2017.
- [54] M. Zolfaghari, K. Singh, and T. Brox. ECO: efficient convolutional network for on-line video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

