Unified Multisensory Perception:
Weakly-Supervised Audio-Visual Video Parsing

Yapeng Tian[1], Dingzeyu Li[2], and Chenliang Xu[1]

[1] University of Rochester
[2] Adobe Research
{yapengtian,chenliang.xu}@rochester.edu; dinli@adobe.com

Abstract. In this paper, we introduce a new problem, named audio-visual video parsing, which aims to parse a video into temporal event segments and label them as either audible, visible, or both. Such a problem is essential for a complete understanding of the scene depicted inside a video. To facilitate exploration, we collect a *Look, Listen, and Parse* (LLP) dataset to investigate audio-visual video parsing in a weakly-supervised manner. This task can be naturally formulated as a Multimodal Multiple Instance Learning (MMIL) problem. Concretely, we propose a novel hybrid attention network to explore unimodal and cross-modal temporal contexts simultaneously. We develop an attentive MMIL pooling method to adaptively explore useful audio and visual content from different temporal extent and modalities. Furthermore, we discover and mitigate modality bias and noisy label issues with an individual-guided learning mechanism and label smoothing technique, respectively. Experimental results show that the challenging audio-visual video parsing can be achieved even with only video-level weak labels. Our proposed framework can effectively leverage unimodal and cross-modal temporal contexts and alleviate modality bias and noisy labels problems.

Keywords: Audio-visual video parsing, weakly-supervised, LLP dataset.

1 Introduction

Human perception involves complex analyses of visual, auditory, tactile, gustatory, olfactory, and other sensory data. Numerous psychological and brain cognitive studies [3,20,46,51] show that combining different sensory data is crucial for human perception. However, the vast majority of work [9,26,48,64] in scene understanding, an essential perception task, focuses on visual-only methods ignoring other sensory modalities. They are inherently limited. For example, when the object of interest is outside of the field-of-view (FoV), one would rely on audio cues for localization. While there is little data on tactile, gustatory, or olfactory signals, we do have an abundance of multimodal audiovisual data, e.g., YouTube videos.

Utilizing and learning from both auditory and visual modalities is an emerging research topic. Recent years have seen progress in learning representations [1,2,19,23,37,38], separating visually indicated sounds [8,11,12,13,65,66,10,70],
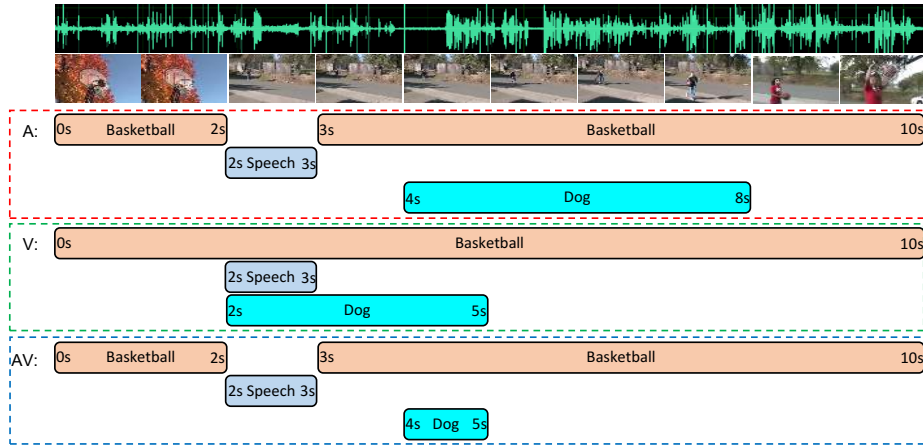
Fig. 1: Our audio-visual video parsing model aims to parse a video into different audio (audible), visual (visible), and audio-visual (audi-visible) events with correct categories and boundaries. A dog in the video visually appears from 2nd second to 5th second and make barking sounds from 4th second to 8th second. So, we have audio event (4s-8s), visual event (2s-5s), and audio-visual event (4s-5s) for the *Dog* event category.

spatially localizing visible sound sources [37,45,55], and temporally localizing audio-visual synchronized segments [27,55,63]. However, past approaches usually assume audio and visual data are always correlated or even temporally aligned. In practice, when we analyze the video scene, many videos have audible sounds, which originate outside of the FoV, leaving no visual correspondences, but still contribute to the overall understanding, such as out-of-screen running cars and a narrating person. Such examples are ubiquitous, which leads us to some basic questions: what video events are audible, visible, and "audi-visible," where and when are these events inside of a video, and how can we effectively detect them?

To answer the above questions, we pose and try to tackle a fundamental problem: *audio-visual video parsing* that recognizes event categories bind to sensory modalities, and meanwhile, finds temporal boundaries of when such an event starts and ends (see Fig. 1). However, learning a fully supervised audio-visual video parsing model requires densely annotated event modality and category labels with corresponding event onsets and offsets, which will make the labeling process extremely expensive and time-consuming. To avoid tedious labeling, we explore weakly-supervised learning for the task, which only requires sparse labeling on the presence or absence of video events. The weak labels are easier to annotate and can be gathered in a large scale from web videos.

We formulate the weakly-supervised audio-visual video parsing as a Multimodal Multiple Instance Learning (MMIL) problem and propose a new framework to solve it. Concretely, we use a new hybrid attention network (HAN) for leveraging unimodal and cross-modal temporal contexts simultaneously. We develop

an attentive MMIL pooling method for adaptively aggregating useful audio and visual content from different temporal extent and modalities. Furthermore, we discover modality bias and noisy label issues and alleviate them with an individual-guided learning mechanism and label smoothing [42], respectively.

To facilitate our investigations, we collect a *Look, listen, and Parse* (LLP) dataset that has $11,849$ YouTube video clips from 25 event categories. We label them with sparse video-level event labels for training. For evaluation, we label a set of precise labels, including event modalities, event categories, and their temporal boundaries. Experimental results show that it is tractable to learn audio-visual video parsing even with video-level weak labels. Our proposed HAN model can effectively leverage multimodal temporal contexts. Furthermore, modality bias and noisy label problems can be addressed with the proposed individual learning strategy and label smoothing, respectively. Besides, we make a discussion on the potential applications enabled by audio-visual video parsing.

The contributions of our work include: (1) a new audio-visual video parsing task towards a unified multisensory perception; (2) a novel hybrid attention network to leverage unimodal and cross-modal temporal contexts simultaneously; (3) an effective attentive MMIL pooling to aggregate multimodal information adaptively; (4) a new individual guided learning approach to mitigate the modality bias in the MMIL problem and label smoothing to alleviate noisy labels; and (5) a newly collected large-scale video dataset, named LLP, for audio-visual video parsing. Dataset, code, and pre-trained models are publicly available in https://github.com/YapengTian/AVVP-ECCV20.

2   Related Work

In this section, we discuss some related work on temporal action localization, sound event detection, and audio-visual learning.

Temporal Action Localization. Temporal action localization (TAL) methods usually use sliding windows as action candidates and address TAL as a classification problem [9,25,29,47,48,67] learning from full supervisions. Recently, weakly-supervised approaches are proposed to solve the TAL. Wang *et al.* [60] present an UntrimmedNet with a classification module and a selection module to learn the action models and reason about the temporal duration of action instances, respectively. Hide-and-seek [49] randomly hides certain sequences while training to force the model to explore more discriminative content. Paul *et al.* [40] introduce a co-activity similarity loss to enforce instances in the same class to be similar in the feature space. Inspired by the class activation map method [68], Nguyen *et al.* [36] propose a sparse temporal pooling network (STPN). Liu *et al.* [28] incorporate both action completeness modeling and action-context separation into a weakly-supervised TAL framework. Unlike actions in TAL, video events in audio-visual video parsing might contain motionless or even out-of-screen sound sources and the events can be perceived by either audio or visual modalities. Even though, we extend two recent weakly-supervised TAL methods:
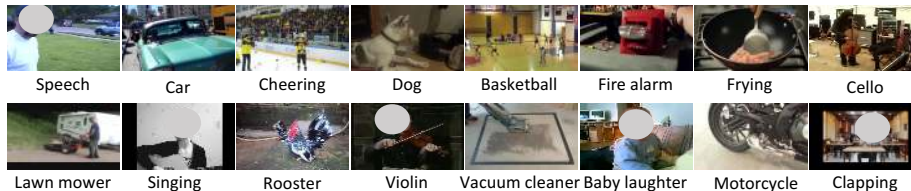
Fig. 2: Some examples from the LLP dataset.

STPN [36] and CMCS [28] to address visual event parsing and compare them with our model in Sec. 6.2.

Sound Event Detection. Sound event detection (SED) is a task of recognizing and locating audio events in acoustic environments. Early supervised approaches rely on some machine learning models, such as support vector machines [7], Gaussian mixture models [17] and recurrent neural networks [39]. To bypass strongly labeled data, weakly-supervised SED methods are developed [6,22,31,62]. These methods only focus on audio events from constrained domains, such as urban sounds [44] and domestic environments [32] and visual information is ignored. However, our audio-visual video parsing will exploit both modalities to parse not only event categories and boundaries but also event perceiving modalities towards a unified multisensory perception for unconstrained videos.

Audio-Visual Learning. Benefiting from the natural synchronization between auditory and visual modalities, audio-visual learning has enabled a set of new problems and applications including representation learning [1,2,19,23,35,37,38], audio-visual sound separation [8,11,12,13,65,66,10,70], vision-infused audio in-painting [69], sound source spatial localization [37,45,55], sound-assisted action recognition [14,21,24], audio-visual video captioning [41,53,54,61], and audio-visual event localization [27,55,56,63]. Most previous work assumes that temporally synchronized audio and visual content are always matched conveying the same semantic meanings. However, unconstrained videos can be very noisy: sound sources might not be visible (*e.g.*, an out-of-screen running car and a narrating person) and not all visible objects are audible (*e.g.*, a static motorcycle and people dancing with music). Different from previous methods, we pose and seek to tackle a fundamental but unexplored problem: audio-visual video parsing for parsing unconstrained videos into a set of video events associated with event categories, boundaries, and modalities. Since the existing methods cannot directly address our problem, we modify the recent weakly-supervised audio-visual event localization methods: AVE [55] and AVSDN [27] adding additional audio and visual parsing branches as baselines.

3   LLP: The Look, Listen and Parse Dataset

To the best of our knowledge, there is no existing dataset that is suitable for our research. Thus, we introduce a *Look, Listen, and Parse* dataset for audio-visual

video scene parsing, which contains 11,849 YouTube video clips spanning over 25 categories for a total of 32.9 hours collected from AudioSet [15]. A wide range of video events (*e.g.*, human speaking, singing, baby crying, dog barking, violin playing, and car running, and vacuum cleaning *etc.*) from diverse domains (*e.g.*, human activities, animal activities, music performances, vehicle sounds, and domestic environments) are included in the dataset. Some examples in the LLP dataset are shown in Fig. 2.

Videos in the LLP have 11,849 video-level event annotations on the presence or absence of different video events for facilitating weakly-supervised learning. Each video is 10s long and has at least 1s audio or visual events. There are 7,202 videos that contain events from more than one event categories and per video has averaged 1.64 different event categories. To evaluate audio-visual scene parsing performance, we annotate individual audio and visual events with second-wise temporal boundaries for randomly selected 1,849 videos from the LLP dataset. Note that the audio-visual event labels can be derived from the audio and visual event labels. Finally, we have totally 6,626 event annotations, including 4,131 audio events and 2,495 visual events for the 1,849 videos. Merging the individual audio and visual labels, we obtain 2,488 audio-visual event annotations. To do validation and testing, we split the subset into a validation set with 649 videos and a testing set with 1,200 videos. Our weakly-supervised audio-visual video parsing network will be trained using the 10,000 videos with weak labels and the trained models are developed and tested on the validation and testing sets with fully annotated labels, respectively.

4 Audio-Visual Video Parsing with Weak Labels

We define the *Audio-Visual Video Parsing* as a task to group video segments and parse a video into different temporal audio, visual, and audio-visual events associated with semantic labels. Since event boundary in the LLP dataset was annotated at second-level, video events will be parsed at scene-level not object/instance level in our experimental setting. Concretely, given a video sequence containing both audio and visual tracks, we divide it into T non-overlapping audio and visual snippet pairs $\{V_t, A_t\}_{t=1}^T$, where each snippet is 1s long and V_t and A_t denote visual and audio content in the same video snippet, respectively. Let $\boldsymbol{y}_t = \{(y_t^a, y_t^v, y_t^{av}) | [y_a^t]_c, [y_v^t]_c, [y_{av}^t]_c \in \{0, 1\}, c = 1, ..., C\}$ be the event label set for the video snippet $\{V_t, A_t\}$, where c refers to the c-th event category and y_t^a, y_t^v, and y_t^{av} denote audio, visual, and audio-visual event labels, respectively. Here, we have a relation: $y_t^{av} = y_t^a * y_t^v$, which means that audio-visual events occur only when there exists both audio and visual events at the same time and from the same event categories.

In this work, we explore the audio-visual video parsing in a weakly-supervised manner. We only have video-level labels for training, but will predict precise event label sets for all video snippets during testing, which makes the weakly-supervised audio-visual video parsing be a multi-modal multiple instance learning (MMIL) problem. Let a video sequence with T audio and visual snippet pairs be a bag.
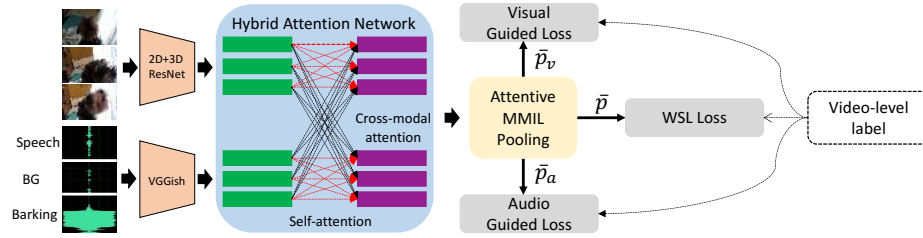
Fig. 3: The proposed audio-visual video parsing framework. It uses pre-trained CNNs to extract snippet-level audio and visual features and leverages multimodal temporal contexts with the proposed hybrid attention network (HAN). For each snippet, we will predict both audio and visual event labels from the aggregated features by the HAN. Attentive MMIL pooling is utilized to adaptively predict video-level event labels for weakly-supervised learning (WSL) and individual guided learning is devised to mitigate the modality bias issue.

Unlike the previous audio-visual event localization [55] that is formulated as a MIL problem [30] where an audio-visual snippet pair is regarded as an instance, each audio snippet and the corresponding visual snippet occurred at the same time denote two individual instances in our MMIL problem. So, a positive bag containing video events will have at least one positive video snippet; meanwhile at least one modality has video events in the positive video snippet. During training, we can only access bag labels. During inference, we need to know not only which video snippets have video events but also which sensory modalities perceive the events. The temporal and multi-modal uncertainty in this MMIL problem makes it very challenging.

5 Method

First, we present the overall framework that formulates the weakly-supervised audio-visual video parsing as an MMIL problem in Sec. 5.1. Built upon this framework, we propose a new multimodal temporal model: hybrid attention network in Sec. 5.2; attentive MMIL pooling in Sec. 5.3; addressing modality bias and noisy label issues in Sec. 5.4.

5.1   Audio-Visual Video Parsing Framework

Our framework, as illustrated in Fig. 3, has three main modules: audio and visual feature extraction, multimodal temporal modeling, and attentive MMIL pooling. Given a video sequence with T audio and visual snippet pairs $\{V_t, A_t\}_{t=1}^T$, we first use pre-trained visual and audio models to extract snippet-level visual features: $\{f_v^t\}_{t=1}^T$ and audio features: $\{f_a^t\}_{t=1}^T$, respectively. Taking extracted audio and visual features as inputs, we use two hybrid attention networks as the multimodal temporal modeling module to leverage unimodal and cross-modal

temporal contexts and obtain updated visual features $\{\hat{f}_v^t\}_{t=1}^T$ and audio features $\{\hat{f}_a^t\}_{t=1}^T$. To predict audio and visual instance-level labels and make use of the video-level weak labels, we address the MMIL problem with a novel attentive MMIL pooling module outputting video-level labels.

5.2 Hybrid Attention Network

Natural videos tend to contain continuous and repetitive rather than isolated audio and visual content. In particular, audio or visual events in a video usually redundantly recur many times inside the video, both within the same modality (unimodal temporal recurrence [34,43]), as well as across different modalities (audio-visual temporal synchronization [23] and asynchrony [59]). The observation suggests us to jointly model the temporal recurrence, co-occurrence, and asynchrony in a unified approach. However, existing audio-visual learning methods [27,55,63] usually ignore the audio-visual temporal asynchrony and explore unimodal temporal recurrence using temporal models (*e.g.*, LSTM [18] and Transformer [58]) and audio-visual temporal synchronization using multimodal fusion (*e.g.*, feature fusion [55] and prediction ensemble [21]) in a isolated way. To simultaneously capture multimodal temporal contexts, we propose a new temporal model: Hybrid Attention Network (HAN), which uses a self-attention network and a cross-attention network to adaptively learn which bimodal and cross-modal snippets to look for each audio or visual snippet, respectively.

At each time step t, a hybrid attention function g in HAN will be learned from audio and visual features: $\{f_a^t, f_v^t\}_{t=1}^T$ to update f_a^t and f_v^t, respectively. The updated audio feature $\hat{f}_a^t$ and visual feature $\hat{f}_v^t$ can be computed as:

$$\hat{f}_a^t = g(f_a^t, \boldsymbol{f}_a, \boldsymbol{f}_v) = f_a^t + g_{sa}(f_a^t, \boldsymbol{f}_a) + g_{ca}(f_a^t, \boldsymbol{f}_v) \ , \tag{1}$$

$$\hat{f}_v^t = g(f_v^t, \boldsymbol{f}_a, \boldsymbol{f}_v) = f_v^t + g_{sa}(f_v^t, \boldsymbol{f}_v) + g_{ca}(f_v^t, \boldsymbol{f}_a) \ , \tag{2}$$

where $\boldsymbol{f}_a = [f_a^1; ...; f_a^T]$ and $\boldsymbol{f}_v = [f_v^1; ...; f_v^T]$; g_{sa} and g_{ca} are self-attention and cross-modal attention functions, respectively; skip-connections can help preserve the identity information from the input sequences. The two attention functions are formulated with the same computation mechanism. With $g_{sa}(f_a^t, \boldsymbol{f}_a)$ and $g_{ca}(f_a^t, \boldsymbol{f}_v)$ as examples, they are defined as:

$$g_{sa}(f_a^t, \boldsymbol{f}_a) = \sum_{t=1}^T w_t^{sa} f_a^t = softmax(\frac{f_a^t \boldsymbol{f}_a^{'}}{\sqrt{d}})\boldsymbol{f}_a \ , \tag{3}$$

$$g_{ca}(f_a^t, \boldsymbol{f}_v) = \sum_{t=1}^T w_t^{ca} f_v^t = softmax(\frac{f_a^t \boldsymbol{f}_v^{'}}{\sqrt{d}})\boldsymbol{f}_v \ , \tag{4}$$

where the scaling factor d is equal to the audio/visual feature dimension and $(\cdot)^{'}$ denotes the transpose operator. Clearly, the self-attention and cross-modal attention functions in HAN will assign large weights to snippets, which are similar to the query snippet containing the same video events within the same modality
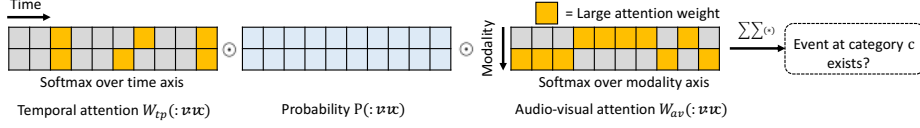
Fig. 4: Attentive MMIL Pooling. For event category c, temporal and audio-visual attention mechanisms will adaptively select informative event predictions crossing temporal and modality axes, respectively, for predicting whether there is an event at the category.

and cross different modalities. The experimental results show that the HAN modeling unimodal temporal recurrence, multimodal temporal co-occurrence, and audio-visual temporal asynchrony can well capture unimodal and cross-modal temporal contexts and improves audio-visual video parsing performance.

5.3   Attentive MMIL Pooling

To achieve audio-visual video parsing, we predict all event labels for audio and visual snippets from temporal aggregated features: $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^T$. We use a shared fully-connected layer to project audio and visual features to different event label space and adopt a sigmoid function to output probability for each event category:

$$p_a^t = sigmoid(FC(\hat{f}_a^t))\ ,\tag{5}$$

$$p_v^t = sigmoid(FC(\hat{f}_v^t))\ ,\tag{6}$$

where p_a^t and p_v^t are predicted audio and visual event probabilities at timestep t, respectively. Here, the shared FC layer can implicitly enforce audio and visual features into a similar latent space. The reason to use sigmoid to output an event probability for each event category rather than softmax to predict a probability distribution over all categories is that a single snippet may have multiple labels rather than only a single event as assumed in Tian *et al.* [55].

Since audio-visual events only occur when sound sources are visible and their sounds are audible, the audio-visual event probability p_{av}^t can be derived from individual audio and visual predictions: $p_{av}^t = p_a^t * p_v^t$. If we have direct supervisions for all audio and visual snippets from different time steps, we can simply learn the audio-visual video parsing network in a fully-supervised manner. However, in this MMIL problem, we can only access a video-level weak label $\bar{y}$ for all audio and visual snippets: $\{A_t, V_t\}_{t=1}^T$ from a video. To learn our network with weak labels, as illustrated in Fig. 4, we propose a attentive MMIL pooling method to predict video-level event probability: $\bar{p}$ from $\{p_a^t, p_v^t\}_{t=1}^T$. Concretely, the $\bar{p}$ is computed by:

$$\bar{p} = \sum_{t=1}^{T} \sum_{m=1}^{M} (W_{tp} \odot W_{av} \odot P)[t, m, :]\ ,\tag{7}$$

where $\odot$ denotes element-wise multiplication; m is a modality index and $M = 2$ refers to audio and visual modalities; W_{tp} and W_{av} are temporal attention and audio-visual attention tensors predicted from $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^T$, respectively, and P is the probability tensor built by $\{p_a^t, p_v^t\}_{t=1}^T$ and we have $P(t, 0, :) = p_a^t$ and $P(t, 1, :) = p_v^t$. To compute the two attention tensors, we first compose an input feature tensor F, where $F(t, 0, :) = \hat{f}_a^t$ and $F(t, 1, :) = \hat{f}_v^t$. Then, two different FC layers are used to transform the F into two tensors: F_{tp} and F_{av}, which has the same size as P. To adpatively select most informative snippets for predicting probabilities of different event categories, we assign different weights to snippets at different time steps with a temporal attention mechanism:

$$W_{tp}[:, m, c] = softmax(F_{tp}[:, m, c]) \ , \tag{8}$$

where $m = 1, 2$ and $c = 1, \ldots, C$. Accordingly, we can adaptively select most informative modalities with the audio-visual attention tensor:

$$W_{av}[t, :, c] = softmax(F_{av}[t, :, c]) \ , \tag{9}$$

where $t = 1, \ldots, T$ and $c = 1, \ldots, C$. The snippets within a video from different temporal steps and different modalities may have different video events. The proposed attentive MMIL pooling can well model this observation with the tensorized temporal and multimodal attention mechanisms.

With the predicted video-level event probability $\bar{p}$ and the ground truth label $\bar{y}$, we can optimize the proposed weakly-supervised learning model with a binary cross-entropy loss function: $\mathcal{L}_{wsl} = CE(\bar{p}, \bar{y}) = \sum_{c=1}^C \bar{y}[c] log(\bar{p}[c])$.

5.4 Alleviating Modality Bias and Label Noise

The weakly supervised audio-visual video parsing framework only uses less detailed annotations without requiring expensive densely labeled audio and visual events for all snippets. This advantage makes this weakly supervised learning framework appealing. However, it usually enforces models to only identify discriminative patterns in the training data, which was observed in previous weakly-supervised MIL problems [49,50,68]. In our MMIL problem, the issue becomes even more complicated since there are multiple modalities and different modalities might not contain equally discriminative information. With weakly-supervised learning, the model tends to only use information from the most discriminative modality but ignore another modality, which can probably achieve good video classification performance but terrible video parsing performance on the events from ignored modality and audio-visual events. Since a video-level label contains all event categories from audio and visual content within the video, to alleviate the modality bias in the MMIL, we propose to use explicit supervisions to both modalities with a guided loss:

$$\mathcal{L}_g = CE(\bar{p}_a, \bar{y}_a) + CE(\bar{p}_v, \bar{y}_v) \ , \tag{10}$$

where $\bar{y}_a = \bar{y}_v = \bar{y}$, and $\bar{p}_a = \sum_{t=1}^T (W_{tp} \odot P)[t, 0, :]$ and $\bar{p}_v = \sum_{t=1}^T (W_{tp} \odot P)[t, 1, :]$ are video-level audio and visual event probabilities, respectively.

However, not all video events are audio-visual events, which means that an event occurred in one modality might not occur in another modality and then the corresponding event label will be label noise for one of the two modalities. Thus, the guided loss: $\mathcal{L}_g$ suffers from noisy label issue. For the example shown in Fig. 3, the video-level label is $\{Speech, Dog\}$ and the video-level visual event label is only $\{Dog\}$. The $\{Speech\}$ will be a noisy label for the visual guided loss. To handle the problem, we use label smoothing [52] to lower the confidence of positive labels with smoothing $\bar{\boldsymbol{y}}$ and generate smoothed labels: $\bar{\boldsymbol{y}}_a$ and $\bar{\boldsymbol{y}}_v$. They are formulated as: $\bar{\boldsymbol{y}}_a = (1 \quad \epsilon_a)\bar{\boldsymbol{y}} + \frac{\epsilon_a}{K}$ and $\bar{\boldsymbol{y}}_v = (1 \quad \epsilon_v)\bar{\boldsymbol{y}} + \frac{\epsilon_v}{K}$, where $\epsilon_a, \epsilon_v \in [0,1)$ are two confidence parameters to balance the event probability distribution and a uniform distribution: $u = \frac{1}{K}$ $(K > 1)$. For a noisy label at event category c, when $\bar{\boldsymbol{y}}[c] = 1$ and real $\bar{\boldsymbol{y}}_a[c] = 0$, we have $\bar{\boldsymbol{y}}[c] = (1 \quad \epsilon_a)\bar{\boldsymbol{y}}[c] + \epsilon_a > (1 \quad \epsilon_a)\bar{\boldsymbol{y}} + \frac{\epsilon_a}{K} = \bar{\boldsymbol{y}}_a[c]$ and the smoothed label will become more reliable. Label smoothing technique is commonly adopted in a lot of tasks, such as image classification [52], speech recognition [5], and machine translation [58] to reduce over-fitting and improve generalization capability of deep models. Different from the past methods, we use smoothed labels to mitigate label noise occurred in the individual guided learning. Our final model is optimized with the two loss terms: $\mathcal{L} = \mathcal{L}_{wsl} + \mathcal{L}_g$.

6   Experiments

6.1   Experimental Settings

Implementation Details. For a 10-second-long video, we first sample video frames at $8fps$ and each video is divided into non-overlapping snippets of the same length with 8 frames in 1 second. Given a visual snippet, we extract a 512-D snippet-level feature with fused features extracted from ResNet152 [16] and 3D ResNet [57]. In our experiments, batch size and number of epochs are set as 16 and 40, respectively. The initial learning rate is $3e$-4 and will drop by multiplying 0.1 after every 10 epochs. Our models optimized by Adam can be trained using one NVIDIA 1080Ti GPU.

Baselines. Since there are no existing methods to address the audio-visual video parsing, we design several baselines based on previous state-of-the-art weakly-supervised sound detection [22,62], temporal action localization [28,36], and audio-visual event localization [27,55] methods to validate the proposed framework. To make [27,55] possible to address audio-visual scene parsing, we add additional audio and visual branches to predict audio and visual event probabilities supervised with an additional guided loss as defined in Sec. 5.4. For fair comparisons, the compared approaches use the same audio and visual features as our method.

Evaluation Metrics. To comprehensively measure the performance of different methods, we evaluate them on parsing all types of events (individual audio, visual, and audio-visual events) under both segment-level and event-level metrics. To evaluate overall audio-visual scene parsing performance, we also compute aggregated results, where Type@AV computes averaged audio, visual, and audio-visual event evaluation results and Event@AV computes the F-score considering all

Table 1: Audio-visual video parsing accuracy (%) of different methods on the LLP test dataset. These methods all use the same audio and visual features as inputs for a fair comparison. The top-1 results in each line are highlighted.

Event type	Methods	Segment-level	Event-level
Audio	Kong *et. al* 2018 [22]	39.6	29.1
	TALNet [62]	50.0	41.7
	AVE [55]	47.2	40.4
	AVSDN [27]	47.8	34.1
	Ours	**60.1**	**51.3**
Visual	STPN [36]	46.5	41.5
	CMCS [28]	48.1	45.1
	AVE [55]	37.1	34.7
	AVSDN [27]	52.0	46.3
	Ours	**52.9**	**48.9**
Audio-Visual	AVE [55]	35.4	31.6
	AVSDN [27]	37.1	26.5
	Ours	**48.9**	**43.0**
Type@AV	AVE [55]	39.9	35.5
	AVSDN [27]	45.7	35.6
	Ours	**54.0**	**47.7**
Event@AV	AVE [55]	41.6	36.5
	AVSDN [27]	50.8	37.7
	Ours	**55.4**	**48.0**

audio and visual events for each sample rather than directly averaging results from different event types as the Type@AV. We use both segment-level and event-level F-scores [33] as metrics. The segment-level metric can evaluate snippet-wise event labeling performance. For computing event-level F-score results, we extract events with concatenating positive consecutive snippets in the same event categories and compute the event-level F-score based on mIoU = 0.5 as the threshold.

6.2   Experimental Comparison

To validate the effectiveness of the proposed audio-visual video parsing network, we compare it with weakly-supervised sound event detection methods: Kong *et al* 2018 [22] and TALNet [62] on audio event parsing, weakly-supervised action localization methods: STPN [36] and CMCS [28] on visual event parsing, and modified audio-visual event localization methods: AVE [55] and AVSD [27] on audio, visual, and audio-visual event parsing. The quantitative results are shown in Tab. 2. We can see that our method outperforms compared approaches on all audio-visual video parsing subtasks under both the segment-level and event-level metrics, which demonstrates that our network can predict more accurate snippet-

Table 2: Audio-visual video parsing accuracy (%) of different methods on the LLP test dataset. These methods all use the same audio and visual features as inputs for a fair comparison. The top-1 results in each line are highlighted.

Event type	Methods	Segment-level	Event-level
Audio	Kong *et. al* 2018	39.6	29.1
	TALNet	50.0	41.7
	AVE	47.2	40.4
	AVSDN	47.8	34.1
	Ours	**60.1**	**51.3**
Visual	STPN	46.5	41.5
	CMCS	48.1	45.1
	AVE	37.1	34.7
	AVSDN	52.0	46.3
	Ours	**52.9**	**48.9**
Audio-Visual	AVE	35.4	31.6
	AVSDN	37.1	26.5
	Ours	**48.9**	**43.0**
Type@AV	AVE	39.9	35.5
	AVSDN	45.7	35.6
	Ours	**54.0**	**47.7**
Event@AV	AVE	41.6	36.5
	AVSDN	50.8	37.7
	Ours	**55.4**	**48.0**

wise event categories with more precise event onsets and offsets for testing videos.

Individual Guided Learning. From Tab. 3, we observe that the model without individual guided learning can achieve pretty good performance on audio event parsing but incredibly bad visual parsing results leading to terrible audio-visual event parsing; w/ only $\mathcal{L}_g$ model can achieve both reasonable audio and visual event parsing results; our model trained with both $\mathcal{L}_{wsl}$ and $\mathcal{L}_g$ outperforms model train without and with only $\mathcal{L}_g$. The results indicate that the model trained only $\mathcal{L}_{wsl}$ find discriminative information from mostly sounds and visual information is not well-explored during training and the individual learning can effectively handle the modality bias issue. In addition, when the network is trained with only $\mathcal{L}_g$, it actually models audio and visual event parsing as two individual MIL problems in which only noisy labels are used. Our MMIL framework can learn from clean weak labels with $\mathcal{L}_{wsl}$ and handle the modality bias with $\mathcal{L}_g$ achieves the best overall audio-visual video parsing performance. Moreover, we would like to note that the modality bias issue is from audio and visual data unbalance in training videos, which are originally from an audio-oriented dataset: AudioSet. Since the issue occurred after just 1 epoch training, it is not over-fitting.

Table 3: Ablation study on learning mechanism, attentive MMIL pooling, hybrid attention network, and handling noisy labels. Segment-level audio-visual video parsing results are shown. The best results for each ablation study are highlighted.

Loss	MMIL Pooling	Temporal Net	Handle Noisy Label	Audio	Visual	Audio-Visual	Type@AV	Event@AV
$\mathcal{L}_{wsl}$	Attentive	×	×	**56.9**	16.4	17.2	30.2	43.3
$\mathcal{L}_{g}$	Attentive	×	×	42.3	43.9	34.5	40.3	42.0
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	×	×	45.1	**51.7**	**35.0**	**44.0**	**48.9**
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Max	×	×	31.6	43.6	22.5	32.6	39.1
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Mean	×	×	40.2	43.2	**35.0**	39.5	39.7
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	×	×	**45.1**	**51.7**	**35.0**	**44.0**	**48.9**
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	×	×	45.1	51.7	35.0	44.0	48.9
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	GRU [4]	×	52.0	49.4	39.0	46.8	51.0
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	Transformer [58]	×	53.4	**53.8**	41.8	49.7	53.3
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	×	**58.4**	52.8	**48.4**	**53.2**	**54.5**
$\mathcal{L}_{wsl}$	Attentive	HAN	×	39.6	40.5	20.1	33.4	44.9
$\mathcal{L}_{g}$	Attentive	HAN	×	57.5	52.5	47.4	52.5	53.8
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	×	**58.4**	**52.8**	**48.4**	**53.2**	**54.5**
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Max	HAN	×	55.7	52.0	**48.6**	52.1	51.8
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Mean	HAN	×	56.0	51.9	46.3	51.4	52.9
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	×	**58.4**	**52.8**	48.4	**53.2**	**54.5**
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	×	58.4	52.8	48.4	53.2	54.5
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	Bootstrap [42]	59.0	52.6	47.8	53.1	55.2
$\mathcal{L}_{wsl}+\mathcal{L}_{g}$	Attentive	HAN	Label Smoothing [52]	**60.1**	**52.9**	**48.9**	**54.0**	**55.4**

Attentive MMIL Pooling. To validate the proposed Attentive MMIL Pooling, we compare it with two commonly used methods: Max pooling and Mean pooling. Our Attentive MMIL Pooling (see Tab. 3) is superior over the both compared methods. The Max MMIL pooling only selects the most discriminative snippet for each training video, thus it cannot make full use of informative audio and visual content. The Mean pooling does not distinguish the importance of different audio and visual snippets and equally aggregates instance scores in a bad way, which can obtain good audio-visual event parsing but poor individual audio and visual event parsing since a lot of audio-only and visual-only events are incorrectly parsed as audio-visual events. Our attentive MMIL pooling allows assigning different weights to audio and visual snippets within a video bag for each event category, thus can adaptively discover useful snippets and modalities.
Hybrid Attention Network. We compare our HAN with two popular temporal networks: GRU and Transformer and a base model without temporal modeling in Tab. 3. The models with GRU and Transformer are better than the base model and our HAN outperforms the GRU and Transformer. The results demonstrate that temporal aggregation with exploiting temporal recurrence is important for audio-visual video parsing and our HAN with jointly exploring unimodal temporal recurrence, multimodal temporal co-occurrence, and audio-visual temporal asynchrony is more effective in leveraging the multimodal temporal contexts. Another surprising finding of the HAN is that it actually tends to alleviate the modality bias by enforcing cross-modal modeling.
Noisy Label. Tab. 3 also shows results of our model without handling the noisy label, with Bootstrap [42] and label smoothing-based method. We can find that

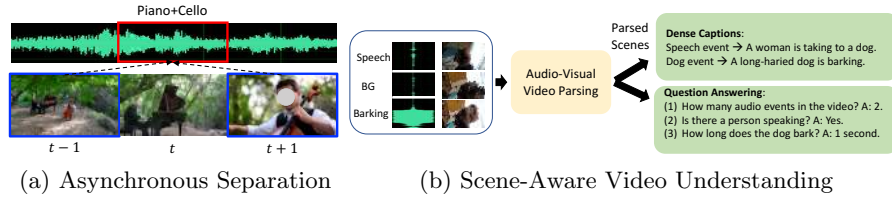(a) Asynchronous Separation   (b) Scene-Aware Video Understanding

Fig. 5: Potential applications of audio-visual video parsing. (a) Temporally asynchronous visual events detected by audio-visual video parsing highlighted in blue boxes can provide related visual information to separate *Cello* sound from the audio mixture in the red box. (b) Parsed scenes can provide important cues for audio-visual scene-aware video dense captioning and question answering.

Bootstrap updating labels using event predictions even decreases performance due to error propagation. Label smoothing-based method with reducing confidence for potential false positive labels can help to learn a more robust model with improved audio-visual video parsing results.

7   Limitation

To mitigate the modality bias issue, the guided loss is introduced to enforce that each modality should also be able to make the correct prediction on its own. Then, a new problem appears: the guide loss is not theoretically correct because some of the events only appear in one modality, so the labels are wrong. Finally, the label smoothing is used to alleviate the label noise. Although the proposed methods work at each step, they also introduce new problems. It is worth to design a one-pass approach. One possible solution is to introduce a new learning strategy to address the modality bias problem rather than using the guided loss. For example, we could perform modality dropout to enforce the model to explore both audio and visual information during training.

8   Conclusion and Future Work

In this work, we investigate a fundamental audio-visual research problem: audio-visual video parsing in a weakly-supervised manner. We introduce baselines and propose novel algorithms to address the problem. Extensive experiments on the newly collected LLP dataset support our findings that the audio-visual video parsing is tractable even learning from cheap weak labels, and the proposed model is capable of leveraging multimodal temporal contexts, dealing with modality bias, and mitigating label noise. Accurate audio-visual video parsing opens the door to a wide spectrum of potential applications, as discussed below.

Asynchronous Audio-Visual Sound Separation. Audio-visual sound separation approaches use sound sources in videos as conditions to separate the

visually indicated individual sounds from sound mixtures [8,11,12,13,65,66]. The underlying assumption is that sound sources are visible. However, sounding objects can be occluded or not recorded in videos and the existing methods will fail to handle these cases. Our audio-visual video parsing model can find temporally asynchronous cross-modal events, which can help to alleviate the problem. For the example in Fig. 5 (a), the existing audio-visual separation models will fail to separate the *Cello* sound from the audio mixture at the time step t, since the sound source *Cello* is not visible in the segment. However, our model can help to find temporally asynchronous visual events with the same semantic label as the audio event *Cello* for separating the sound. In this way, we can improve the robustness of audio-visual sound separation by leveraging temporally asynchronous visual content identified by our audio-visual video parsing models.

Audio-Visual Scene-Aware Video Understanding. The current video understanding community usually focuses on the visual modality and regards information from sounds as a bonus assuming that audio content should be associated with the corresponding visual content. However, we want to argue that auditory and visual modalities are equally important and most natural videos contain numerous audio, visual, and audio-visual events rather than only visual and audio-visual events. Our audio-visual scene parsing can achieve a unified multisensory perception, therefore it has the potential to help us build an audio-visual scene-aware video understanding system regarding all audio and visual events in videos(see Figure 5 (b)).

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017) 1, 4
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. pp. 892–900 (2016) 1, 4
3. Bulkin, D.A., Groh, J.M.: Seeing sounds: visual and auditory interactions in the brain. Current opinion in neurobiology **16**(4), 415–419 (2006) 1
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014) 13
5. Chorowski, J., Jaitly, N.: Towards better decoding and language model integration in sequence to sequence models. arXiv preprint arXiv:1612.02695 (2016) 10
6. Chou, S.Y., Jang, J.S.R., Yang, Y.H.: Learning to recognize transient sound events using attentional supervision. In: IJCAI. pp. 3336–3342 (2018) 4
7. Elizalde, B., Kumar, A., Shah, A., Badlani, R., Vincent, E., Raj, B., Lane, I.: Experiments on the dcase challenge 2016: Acoustic scene classification and sound event detection in real life recording. arXiv preprint arXiv:1607.06706 (2016) 4
8. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018) 1, 4, 15
9. Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. IEEE transactions on pattern analysis and machine intelligence **35**(11), 2782–2795 (2013) 1, 3
10. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10478–10487 (2020) 1, 4
11. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 35–53 (2018) 1, 4, 15
12. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019) 1, 4, 15
13. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3879–3888 (2019) 1, 4, 15
14. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. arXiv preprint arXiv:1912.04487 (2019) 4
15. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017) 5
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 10
17. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing **2013**(1), 1 (2013) 4
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 7

19. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9248–9257 (2019) 1, 4
20. Jacobs, R.A., Xu, C.: Can multisensory training aid visual learning? a computational investigation. Journal of vision **19**(11), 1–1 (2019) 1
21. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5492–5501 (2019) 4, 7
22. Kong, Q., Xu, Y., Wang, W., Plumbley, M.D.: Audio set classification with attention model: A probabilistic perspective. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 316–320. IEEE (2018) 4, 10, 11
23. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Advances in Neural Information Processing Systems. pp. 7763–7774 (2018) 1, 4, 7
24. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6232–6242 (2019) 4
25. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018) 3
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1
27. Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006. IEEE (2019) 2, 4, 7, 10, 11
28. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019) 3, 4, 10, 11
29. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 344–353 (2019) 3
30. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Advances in neural information processing systems. pp. 570–576 (1998) 6
31. McFee, B., Salamon, J., Bello, J.P.: Adaptive pooling operators for weakly labeled sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(11), 2180–2193 (2018) 4
32. Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., Virtanen, T.: Dcase 2017 challenge setup: Tasks, datasets and baseline system (2017) 4
33. Mesaros, A., Heittola, T., Virtanen, T.: Metrics for polyphonic sound event detection. Applied Sciences **6**(6), 162 (2016) 11
34. Naphade, M.R., Huang, T.S.: Discovering recurrent events in video using unsupervised methods. In: Proceedings. International Conference on Image Processing. vol. 2, pp. II–II. IEEE (2002) 7
35. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011) 4

36. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6752–6761 (2018) 3, 4, 10, 11

37. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018) 1, 2, 4

38. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European Conference on Computer Vision. pp. 801–816 (2016) 1, 4

39. Parascandolo, G., Huttunen, H., Virtanen, T.: Recurrent neural networks for polyphonic sound event detection in real life recordings. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6440–6444. IEEE (2016) 4

40. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 563–579 (2018) 3

41. Rahman, T., Xu, B., Sigal, L.: Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8908–8917 (2019) 4

42. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014) 3, 13

43. Roma, G., Nogueira, W., Herrera, P., de Boronat, R.: Recurrence quantification analysis features for auditory scene classification. IEEE AASP challenge on detection and classification of acoustic scenes and events 2 (2013) 7

44. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1041–1044 (2014) 4

45. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4358–4366 (2018) 2, 4

46. Shams, L., Seitz, A.R.: Benefits of multisensory learning. Trends in cognitive sciences 12(11), 411–417 (2008) 1

47. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5734–5743 (2017) 3

48. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1049–1058 (2016) 1, 3

49. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE international conference on computer vision (ICCV). pp. 3544–3553. IEEE (2017) 3, 9

50. Song, H.O., Lee, Y.J., Jegelka, S., Darrell, T.: Weakly-supervised discovery of visual pattern configurations. In: Advances in Neural Information Processing Systems. pp. 1637–1645 (2014) 9

51. Spence, C., Squire, S.: Multisensory integration: maintaining the perception of synchrony. Current Biology 13(13), R519–R521 (2003) 1

52. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 10, 13

53. Tian, Y., Guan, C., Goodman, J., Moore, M., Xu, C.: An attempt towards interpretable audio-visual video captioning. arXiv preprint arXiv:1812.02872 (2018) 4
54. Tian, Y., Guan, C., Justin, G., Moore, M., Xu, C.: Audio-visual interpretable and controllable video captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019) 4
55. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV. pp. 247–263 (2018) 2, 4, 6, 7, 8, 10, 11
56. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in the wild. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops (2019) 4
57. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018) 10
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 7, 10, 13
59. Vroomen, J., Keetels, M., De Gelder, B., Bertelson, P.: Recalibration of temporal order perception by exposure to audio-visual asynchrony. Cognitive brain research **22**(1), 32–35 (2004) 7
60. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017) 3
61. Wang, X., Wang, Y.F., Wang, W.Y.: Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. arXiv preprint arXiv:1804.05448 (2018) 4
62. Wang, Y., Li, J., Metze, F.: A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 31–35. IEEE (2019) 4, 10, 11
63. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 2, 4, 7
64. Yang, Y., Liu, J., Shah, M.: Video scene understanding using multi-scale analysis. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1669–1676. IEEE (2009) 1
65. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1735–1744 (2019) 1, 4, 15
66. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 570–586 (2018) 1, 4, 15
67. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923 (2017) 3
68. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016) 3, 9
69. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 283–292 (2019) 4

70. Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z.: Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 1, 4