Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception

Junyu Gao[1,2], Mengyuan Chen[1,2], and Changsheng Xu[1,2,3]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Peng Cheng Laboratory, ShenZhen, China
{junyu.gao, csxu}@nlpr.ia.ac.cn; chenmengyuan2021@ia.ac.cn

Abstract

With only video-level event labels, this paper targets at the task of weakly-supervised audio-visual event perception (WS-AVEP), which aims to temporally localize and categorize events belonging to each modality. Despite the recent progress, most existing approaches either ignore the unsynchronized property of audio-visual tracks or discount the complementary modality for explicit enhancement. We argue that, for an event residing in one modality, the modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal. To this end, we propose to collect Cross-Modal Presence-Absence Evidence (CMPAE) in a unified framework. Specifically, by leveraging uni-modal and cross-modal representations, a presence-absence evidence collector (PAEC) is designed under Subjective Logic theory. To learn the evidence in a reliable range, we propose a joint-modal mutual learning (JML) process, which calibrates the evidence of diverse audible, visible, and audi-visible events adaptively and dynamically. Extensive experiments show that our method surpasses state-of-the-arts (e.g., absolute gains of 3.6% *and* 6.1% *in terms of event-level visual and audio metrics). Code is available in* github.com/MengyuanChen21/CVPR2023-CMPAE.

1. Introduction

Research in computer vision places a significant emphasis on the visual aspects of event perception; nevertheless, in the real world with multisensory modalities, natural events are distinguished by a great deal more than just their appearance [11, 30, 52, 53, 56, 66]. For instance, think of playing a specific musical instrument in a concert hall, a barking dog, or starting a car with the engine sound. To properly comprehend an event, it is necessary to take acoustics into account and engage in joint audio-visual perception.

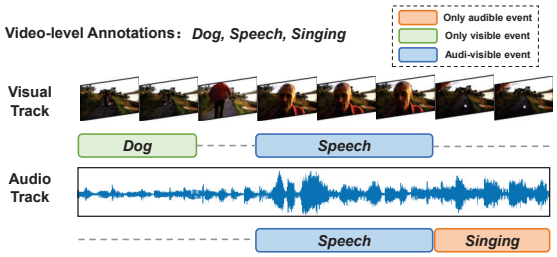

Figure 1. With only video-level annotations, weakly-supervised audio-visual event perception (WS-AVEP) aims to predict the temporal boundaries of various only audible (in orange), only visible (in green), or audi-visible (in blue) events in a video.

The target of audio-visual event perception (AVEP) is to temporally categorize video events. However, collecting precisely temporal audio-visual annotations is a bottleneck and consequently limits the scalability of a fully-supervised learning framework. As a result, Tian *et al*. [52, 53] propose to perceive audio-visual events in an weakly-supervised manner, where only easily available video-level labels are needed during model training. As depicted in Figure 1, given videos which may have various audible, visible, or audi-visible events, the weakly-supervised audio-visual event perception (WS-AVEP) is commonly optimized by utilizing the video-level annotations.

To date in the literature, current WS-AVEP approaches mainly embrace two types of pipelines: (1) To comprehensively incorporate both modalities, some pioneering methods [53] assume that each event in a video is simultaneously audible and visible. Based on this characteristic, numerous cross-modal fusion strategies are proposed, including cross attention [60, 61, 63] and modality interaction [47, 62]. Although achieving promising performance, the rigorous assumption may not always hold in practice

Table 1. Comparison with the state-of-the-art methods on two tasks, AVVP and AVE. Note that the two tasks have different goals and properties. Please refer to the text for more details.

Task \ Method	CMBS [61]	JoMoLD [6]	**Ours**
AVVP [52]	51.7	**57.3**	**60.1**
AVE [53]	**74.2**	71.8	**74.8**

due to some audio-visual non-correspondence caused by out-of-screen objects and background noises. To this end, targeting at unsynchronized audio and visual information modeling, (2) Tian *et al*. [52] suggest a more general setting that recognizes event categories and temporal boundaries bind to sensory modalities, which breaks the modality consistency restriction. Since video-level labels do not indicate the detailed modality information, further research focuses on mining audio- or visual-specific information by learning from modality-specific noises [6], heterogeneous information [58], or hierarchical features [22]. Nonetheless, these approaches discount the complementary modality for explicitly enhancing the prediction of the other modality. Although the multimodal multiple instance learning (MMIL) framework [6, 30, 52] can perform cross-modal enhancement for the feature learning, it still neglects the explicit and extra assistance of the complementary clues for individual modality prediction. Consequently, as shown in Table 1, state-of-the-arts of the two pipelines can only achieve significant performance in one single WS-AVEP setting, showing that current methods are in a dilemma of making full use of both uni-modal and cross-modal information.

To tackle the above issues, we argue that, for an event residing in one modality[1], *the modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal*. On the one hand, to fully tap the potential of each modality, it is desirable to make the modality self-reliable for determining the evidence strength of an present event in the corresponding track. On the other hand, for judging which events are absent, relying on a single modality is insufficient, whereas the other track can hand over complementary but not dominant assistance [66]. For example, although a baby is out-of-screen and the event "*baby_cry*" only appears in the audio modality, we can still infer that the audio track might not contain outdoor events because the perceived visual scene is considered to be indoors. Similarly, when the audio track is salient, some vigorous activity may be less likely to occur in the visual track.

Motivated by the above observations, we aim to capture the presence and absence evidence for individual events by using uni-modal and cross-modal information. To obtain reliable evidence that can explicitly reflect and measure the event presence/absence intensity in each modality, conventional convolutional neural networks, which are based on classification probability, could be overconfident and in the cart [43, 50, 55]. Recently, evidential deep learning (EDL) [36, 50], which can quantify uncertainty in model predictions trustfully by collecting subjective evidence, has attracted increasing attention and been successfully used in a variety of computer vision tasks [1, 3, 5, 17, 27, 57]. In this paper, we propose to collect Cross-Modal Presence-Absence Evidence (CMPAE) for WS-AVEP in a unified framework. As shown in Figure 2, we design a presence-absence evidence collector (PAEC) by using uni-modal and cross-modal representations. Here, the presence evidence of events in each track is derived from the modality itself, whereas the other modality acts as a cross-modal selector for generating the absence evidence. The evidence of each temporal snippet is then accumulated to video-level evidence and optimized in accordance with Subjective Logic theory [23, 64]. To learn the evidence in a reliable range, we propose a joint-modal mutual learning (JML) process, which calibrates the evidence of diverse audible, visible, and audi-visible events adaptively and dynamically. By virtue of the above design, the proposed PAEC and JML modules can cooperate with each other in a unified framework for effective presence-absence evidence learning.

Our main contributions can be summarized as follows:

- We propose a novel cross-modal presence-absence evidence learning framework for weakly-supervised audio-visual event perception, which jointly enjoys the merits of uni-modal discrimination and cross-modal enhancement under Subjective Logic theory.
- With the cooperative presence-absence evidence collector and the joint-modal mutual learning process, we inject the uni-modal and cross-modal information into the learned evidence and calibrate it to a reliable range.
- We conduct extensive and in-depth experiments on several popular and standard WS-AVEP datasets [52, 53]. The encouraging results compared with state-of-the-arts demonstrate the effectiveness of our method.

2. Related Work

Audio-Visual Learning. Living in the multi-modal world with fruitful audio and visual information, humans understand events via seeing and hearing from the environments [8–10, 12, 30, 52, 66]. To perceive the world both visually and aurally, learning audio-visual representation simultaneously is fundamental [2, 14, 20, 34, 40, 41, 44]. To obtain effective multi-modal representation, cross-modal attention mechanisms [26, 30, 60, 66] are commonly leveraged for audio-visual feature fusion. With the joint multi-modal representation, several tasks are explored, such as vision-infused audio inpainting [69], sound-assisted action recognition [24, 25], source sound localization/separation [13,

[1]No matter whether the event is modality-specific or audi-visible.

49], zero-shot learning [37], question answering [28], multi-modal video domain adaptation [42, 66], and audio-visual video parsing/localization [52, 53].

Weakly-supervised Audio-Visual Event Perception. To comprehensively leverage both audio and visual modalities to understand video in a weakly-supervised manner, Tian *et al*. firstly introduce the audio-visual event (AVE) localization task [53]. In the AVE task, when an event is both auditory and visible at the same time, the model determines its presence and pinpoints its border in the temporal dimension. Existing approaches frequently rely on attention strategies [31, 32, 59, 60, 63] or cross-modal interactions [29, 35, 46–48, 62] to acquire effective representations for AVE. Other works adopt additional regular terms to improve the discriminative ability of models, such as background suppression [61] and positive sample propagation [70]. Different from AVE, another task named audio-visual video parsing (AVVP) [52] disproves the assumption that audio and visual signals are always temporally synchronized and in alignment. Based on the hybrid attention network and multi-modal multiple instance learning (MMIL) framework [52], various strategies [6, 22, 30, 38, 39, 58] are explored, such as audio-visual track swapping and contrasting [58], cross-video and cross-modality enhancement [30], dual hierarchical hybrid network [22], and joint-modal label denoising [6]. JoMoLD [6] dynamically identifies and removes modality-specific noisy labels in a two-stage manner. Despite their significant performance, the above methods discount the complementary modality information for explicitly enhancing and calibrating the prediction of individual modalities. Currently, most methods target at either AVE or AVVP separately due to the different properties between the two tasks. Although a few recent methods [65, 71] attempt to conduct experiments on both tasks, they still employ different baseline frameworks severally for each dataset. In this paper, we propose a unified CMPAE framework that can handle both AVE and AVVP tasks.

Evidential Deep Learning (EDL). The mainstream deep networks essentially perform a point estimation of the classification probability distribution, which cannot quantify the predictive uncertainty and have a tendency to be overconfident in false predictions [16]. To this end, EDL [36, 50] targets at knowing "what they don't know" and falling back onto a prior belief. Based on the Dempster-Shafer theory of evidence [64] and Subjective Logic theory [23], EDL allows uncertainty estimation in a single forward pass [55] by collecting evidence for each category and modeling the distribution of class probabilities. In recent two years, EDL has received increasing attention and has successfully been adopted in various computer vision tasks, including multi-view classification [17, 33], open-set recognition and out-of-distribution detection [3, 21], regression [1], long-tail learning [27], meta-learning [45], *etc*. Some pioneering works also explore the temporal localization and weakly-supervised tasks in videos [5, 57]. However, the above approaches neglect the joint-modal learning in presence-absence evidence collection and calibration.

3. Our Approach

Our proposed CMPAE framework is shown in Figure 2, given a video containing audio and visual tracks, we first utilize pre-trained feature extractors to obtain cross-modal features of each snippet (Section 3.1). Then, under the evidential deep learning framework, a presence-absence evidence collector (PAEC) for each modality is designed, where the presence evidence is obtained by the uni-modal information and the absence evidence is additionally constructed via a cross-modal selector (Section 3.2). The learned evidence is further calibrated to a reliable range by leveraging a joint-modal mutual learning (JML) process adaptively and dynamically (Section 3.3). Finally, the unified framework is end-to-end learned (Section 3.4).

3.1. Notations and Preliminaries

The WS-AVEP task targets at localizing audible/visible events that occur in each snippet of a video. Specifically, for a video V, its corresponding multi-hot event category labels are $\mathbf{y}^a, \mathbf{y}^v$ and $\mathbf{y}^{av}$, which denote audio, visual, and audio-visual event labels, respectively. An audio-visual event means that the event appears in both audio and visual tracks in a synchronized fashion. Note that $\mathbf{y}^a, \mathbf{y}^v, \mathbf{y}^{av} \in \{0,1\}^C$, where C is the event category number. However, due to the weakly-supervised setting, we can only access the modality-agnostic video-level label $\mathbf{y} \in \{0,1\}^C$ during training. Following previous works [6, 7, 30, 52, 58], we first divide the video V into T non-overlapping snippets, and use pre-trained off-the-shelf networks and embedding layers, to extract audio and visual features $\{\mathbf{x}_t^a, \mathbf{x}_t^v\}_{t=1}^T$, where the feature dimension of each modality is D for further uni-modal and cross-modal learning.

Currently, existing dominant approaches mainly embrace a video-level classification framework, which learns importance scores for aggregating snippet-level predictions into a video-level one and then performs optimization by using the standard binary cross-entropy (BCE) loss:

$$\mathcal{L}_{bce} = -\sum_{m\in\mathcal{M}}\sum_{c=1}^{C} y_c^m \log p_{vid,c}^m \tag{1}$$

where $\mathcal{M} = \{a, v, av\}$ denotes the set of different tracks. $p_{vid,c}^m$ is the aggregated video-level prediction, which is learned by using the attention mechanism [53, 61] or MMIL formulation [6, 52]. Note that since only video-level annotations are available, some methods [30, 52, 53] treat the labels of the audio, visual, and audio-visual tracks as the same, *i.e*., $\mathbf{y}^a = \mathbf{y}^v = \mathbf{y}^{av} = \mathbf{y}$, which may hinder the optimization of specific modalities. To improve the learning

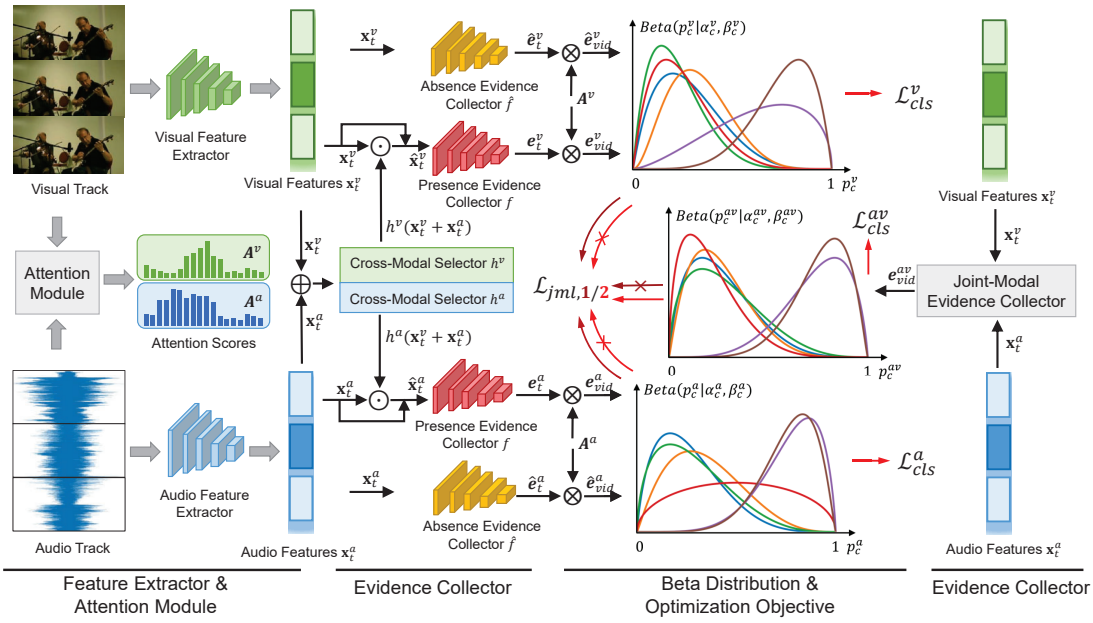

Figure 2. Overall framework of the proposed cross-modal presence-absence evidence learning (CMPAE). Given a video containing audio and visual tracks, we first extract snippet-level visual and audio features. Then, a presence-absence evidence collector (PAEC) for each modality is designed, which collects the presence evidence from the uni-modal information and additionally constructs the absence evidence via cross-modal selectors. In the process of obtaining video-level evidence, an attention module is adopted to generate aggregation weights. Finally, the learned evidence is adaptively and dynamically calibrated by a joint-modal mutual learning (JML) strategy.

quality of different tracks, some methods [6, 58] attempt to mine modality-specific labels for denoising $\mathbf{y}^a$ and $\mathbf{y}^v$.

3.2. Presence-Absence Evidence Collector

We design a novel cross-modal presence-absence evidence collector for the WS-AVEP task by leveraging the formalism of evidential deep learning (EDL) [36, 50] based on Subjective Logic theory [23]. Different from standard classifier learning where the resultant model is ignorant of the confidence of its prediction, EDL proposes to explicitly collect evidence in an uncertainty-aware manner by treating classification output as the pointwise estimation of the categorical distribution and placing a prior over the distribution of all possible classification outputs. It is obvious that, for WS-AVEP, collecting explicit evidence and building an uncertainty-aware framework is even more crucial than it in the traditional single-modality perception tasks [3,5,17,27]: (1) The lack of modality-specific labels leads to significant noise and uncertainty in model optimization. It is difficult to determine the contribution of an individual modality to the final prediction. (2) Due to the lack of fine-grained temporal annotation, the temporally accumulated classification probability could be unreliable thus hinders the uni-modal and cross-modal feature learning.

Based on the above analysis, the traditional BCE loss is incapable of explicitly collecting presence/absence evidence in an uncertainty-aware manner. To this end, some pioneering EDL-based models [50, 68] jointly generate presence and absence evidence from the same single-modality features. However, these approaches do not take advantage of the other complementary modality to improve the prediction of the current one. We argue that, for an event residing in one modality, the modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal. As shown in Figure 2, for the features of modality $m \in \{a, v\}$, $\{\mathbf{x}_t^m\}_{t=1}^T$, the presence evidence of the c-th event category is a scalar and can be obtained by:

$$e_{t,c}^m = g\left(f_c(\mathbf{x}_t^m; \boldsymbol{\theta}_1)\right), \tag{2}$$

where f_c is a DNN parameterized by $\boldsymbol{\theta}_1$ to collect evidence for the c-th event category, g denotes the evidence function, *e.g.*, SoftPlus or Exp, to keep the obtained evidence $e_{t,c}^m$ non-negative. As for the absence evidence, since the complementary modality can provide useful information, we design a cross-modal selector to mine absence-relevant context for improving evidence learning:

$$\begin{aligned} \hat{e}_{t,c}^m &= g\left(\hat{f}_c(\hat{\mathbf{x}}_t^m; \boldsymbol{\theta}_2)\right), \\ \hat{\mathbf{x}}_t^m &= \mathbf{x}_t^m \odot \left(h^m(\mathbf{x}_t^m + \mathbf{x}_t^{\hat{m}}; \boldsymbol{\theta}_3^m) + \mathbf{1}\right), \end{aligned} \tag{3}$$

where $\hat{m}$ denotes the complementary modality of m, and $\mathbf{x}_t^m + \mathbf{x}_t^{\hat{m}}$ fuses the cross-modal features. We assume that each channel of $\mathbf{x}_t^m$ encodes distinct context for absence evidence collection. $h^m(\cdot)$ is a fully-connected layer transforming the fused cross-modality feature into channel-wise

selectors, and $\odot$ is the Hadamard product. $\mathbf{1}$ is an all-ones vector for residual connection. With the snippet-level evidence, the video-level evidence of each modality can be accumulated as: $\{e^m_{vid,c}, \hat{e}^m_{vid,c}\} = \sum_t A^m_{t,c}\{e^m_{t,c}, \hat{e}^m_{t,c}\}$, where $A^m_{t,c}$ is the modality-aware temporal attention score, which can be learned by standard strategies [6, 52]. Note that the evidence collector f_c and $\hat{f}_c$ are shared between audio and visual tracks, which can facilitate the learning of the cross-modal common space and alleviate overfitting issues.

With the video-level presence-absence evidence, under the Subjective Logic theory [23], we build a Beta distributions for the c-th binary classification task in modality m:

$$\text{Beta}(p_c|\alpha_c, \beta_c) = \frac{1}{B(\alpha_c, \beta_c)} p_c^{\alpha_c - 1}(1 - p_c)^{\beta_c - 1}, \quad (4)$$

where $B(\alpha_c, \beta_c) = \Gamma(\alpha_c)\Gamma(\beta_c)/\Gamma(\alpha_c + \beta_c)$ and $\Gamma(\cdot)$ is the Gamma function. For brevity, we omit the scripts m and vid of p, α and β. According to [23], α and β have a fixed relation with the corresponding presence/absence evidence, *i.e.*, $\alpha_c = e_{vid,c} + 1$, $\beta_c = \hat{e}_{vid,c} + 1$. Treating $\text{Beta}(p_c|\alpha_c, \beta_c)$ as the class probability distribution, the Bayes risk for the cross-entropy loss for modality m can be derived as:

$$\begin{aligned}\mathcal{L}^m_{cls} &= \int \left[\sum_{c=1}^{C} -y^m_c \log(p_c)\right] \text{Beta}(p_c|\alpha_c, \beta_c) dp \\ &= \sum_{c=1}^{C} \left[\psi\left(\alpha_c + \beta_c\right) - \psi\left(y^m_c \alpha_c + (1 - y^m_c)\,\beta_c\right)\right],\end{aligned} \quad (5)$$

where $\psi(\cdot)$ is the digamma function.

3.3. Joint-modal Mutual Learning

Although the presence-absence evidence collector leverages the complementary modality for absence-event perception, it primarily focuses on performing single-track recognition with uni-modal information. Nevertheless, the WS-AVEP itself is a cross-modal learning task that involves audio-visual collaboration. As a result, we propose to further learn adaptive and cooperative evidence by performing joint-modal mutual learning between cross-modal evidence and uni-modal evidence.

To perform joint-modal mutual learning effectively and comprehensively, as shown in Figure 2, we first generate global joint-modal presence-absence evidence by using the fused audio-visual features:

$$e^{av}_{vid,c}, \hat{e}^{av}_{vid,c} = \sum_t A^{av}_{t,c} \cdot g(f^{av}_c(\mathbf{x}^a_t + \mathbf{x}^v_t; \boldsymbol{\theta}_4)), \quad (6)$$

where $A^{av}_{t,c}$ is the joint-modal temporal score, which can be obtained by taking the average of A^a_c and A^m_c, and f^{av}_c is a DNN parameterized by $\boldsymbol{\theta}_4$ to collect the global presence-absence evidence of the entire video by using both audio and visual features. To learn the evidence, the EDL loss $\mathcal{L}^{av}_{cls}$ is adopted as in Eq. (5) by using the annotated video-level labels. Note that, different from the evidence collection in each single modality, the global presence and absence evidence are obtained from the same feature due to the information completeness. The detailed analysis about evidence collection strategies can be found in Section 4.3.

After obtaining the presence-absence evidence of the aforementioned three branches, *i.e.* audio, visual, and joint-modal, we design a mutual learning strategy between joint-modal evidence and uni-modal evidence with uncertainty calibration, thus generating more adaptive and comprehensive evidence. According to the Subjective Logic theory [23], the classification probabilities and predictive uncertainties can be inferred as:

$$p^m_c = \frac{e^m_c + 1}{e^m_c + \hat{e}^m_c + 2}, \quad u^m_c = \frac{2}{e^m_c + \hat{e}^m_c + 2}, \quad (7)$$

where $m \in \{a, v, av\}$, and the subscript vid is omitted for brevity. However, since we cannot determine which single modality the labeled event belongs to (or whether it occurs in both modalities simultaniously), it is unreasonable to directly allow the cross-modal classification results to guide the uni-modal learning. As a result, we use the Max operator to fuse the prediction of the target classes on the audio and visual modalities, and the Mean operator for the non-target categories, based on the fact that $y_c = 1$ represents the event c occurred in at least one modality and $y_c = 0$ means that the event c did not occur in either modality. The fusion process can be specifically expressed as follows:

$$\{u^{uni}_c, p^{uni}_c\} = \delta(c)\{u^a_c, p^a_c\} + (1 - \delta(c))\{u^v_c, p^v_c\}, \quad (8)$$

where $\delta(c)$ plays the role of uni-modal selection:

$$\delta(c) = \begin{cases} 1, & p^a_c > p^v_c, y_c = 1, \\ 0, & p^a_c \leq p^v_c, y_c = 1, \\ 1/2, & y_c = 0. \end{cases} \quad (9)$$

Thereafter, we adopt the predictive uncertainties as calibration factors to optimize the joint-modal mutual learning. For each class, since u^{uni}_c reflects the prediction confidence of event c in the most representative modality, we incorporate it to the mutual learning to urge the model to preferentially focus on the corresponding modality and category which are more reliable to learn. In addition, the supervision of the joint-modal branch is error-free in comparison to each individual modality, producing a more stable prediction. Therefore, we aim to make the the joint-modal information dominate the mutual learning when its uncertainty is low. Finally, the optimization objectives of our joint-modal

mutual learning strategy can be expressed as follows:

$$\mathcal{L}_{jml,1} = \sum_{c} (1 - u^{av}) \left(1 - u_c^{uni}\right) * l\left(s(p_c^{av}), p_c^{uni}\right),$$
$$\mathcal{L}_{jml,2} = \sum_{c} u^{av} \left(1 - u_c^{uni}\right) * l\left(p_c^{av}, s(p_c^{uni})\right), \tag{10}$$

where $s(\cdot)$ denotes the gradient truncation operation on the input, and $l(\cdot)$ is a distance metric function, *e.g.* L2-norm. u^{av} is the averaged value of all the u_c^{av} for representing the global uncertainty of the video.

3.4. Learning and Inference

Training. Combining all the aforementioned optimization objectives, we obtain the final loss functions as:

$$\mathcal{L}_i = \sum_{\mathcal{M}}^{m} \mathcal{L}_{cls}^m + \mathcal{L}_{jml,i}, \quad i = 1, 2, \tag{11}$$

where $\mathcal{M} = \{a, v, av\}$. We alternate the loss functions $\mathcal{L}_i$ between iterations to implement the joint-modal mutual learning process. Specifically, the optimization details are summarized in the Supplementary Material.

Inference. For a test video, we first predict its video-level classification probabilities $p_{vid,c}^m$ and the snippet-level temporal class activation sequence $p_{t,c}^m$, where $m \in \{a, v, av\}$. According to Subjective Logic theory, $p_{vid,c}^m = (e_{vid,c}^m + 1)/(e_{vid,c}^m + \hat{e}_{vid,c}^m + 2)$, and $p_{t,c}$ can be inferred similarly by using $e_{t,c}^m$ and $\hat{e}_{t,c}^m$. Thereafter, following the standard process [6, 52], we apply a threshold strategy to obtain proposals for audio and visual events, and finally localize the audio-visual events by taking the intersection of audio and visual events belonging to the same category.

4. Experimental Results

We evaluate CMPAE on two benchmarks: AVVP [52], and AVE [53]. To comprehensively analyze our method for WS-AVEP, we additionally combine LLP and AVE as an entire dataset, named AVEP, for evaluation.

4.1. Experimental Setup

AVVP. Tian *et al.* [52] propose the audio-visual video parsing (AVVP) task and construct the *Look, Listen, and Parse* (LLP) dataset. LLP contains 11,849 10-second video clips of 25 event classes collected from AudioSet [15], and each video contains events of 1.64 categories in average. Since the videos do not guarantee the temporal or categorical consistency of events on the visual and audio tracks, it is challenging and suitable to perform AVVP on this benchmark. Following previous works [6, 30, 52], we split the dataset into a training set of 10,000 videos, a validation set of 649 videos, and a testing set of 1,200 videos.

AVE. The *Audio-Visual Event* (AVE) dataset [53] selects 4,143 YouTube videos covering 28 categories from AudioSet [15]. Different from LLP, AVE focuses on the cases where videos keeps synchronization on both audio and visual tracks, which is the most common situation in the real world. The sizes of the training, validation, and testing sets are 3,339, 402 and 402, respectively.

AVEP. To further evaluate the capacity of different approaches, we additionally combine the LLP and AVE datasets as an entire dataset named AVEP. Specifically, we enlarge the LLP dataset by adding all the AVE samples belonging to categories that are semantically non-overlapping with those in LLP. Ultimately, the AVEP dataset consists of 11,581 training videos, 840 validation videos, and 1,391 testing videos, which cover 39 categories. Note that although the goal and the evaluation metrics of AVEP are the same as those of AVVP, the newly combined dataset is still meaningful: In the dataset of AVVP, most event classes have audible, visible, and audi-visible segments, while all the event categories annotated in AVE are only audi-visible. Therefore, adding new categories with only synchronized information can increase the diversity and difficulty of modality-aware event perception.

Evaluation Metrics. Following previous works [6, 52], we evaluate the predicted audio, visual, and audio-visual event proposals under segment-level and event-level metrics, for both AVVP and AVEP datasets. Specifically, the segment-level metrics include **(1)** F-socre of audio events, **(2)** F-score of visual events, **(3)** F-score of audio-visual events, **(4)** the average of the former three metrics, and **(5)** F-score of all events without considering the modality, whose abbreviations are A, V, AV, Type, and Event, respectively. The event-level metrics are similar, while the event-level F-scores are calculated with a mIoU (mean Intersection over Union) threshold of 0.5. For the AVE task, we follow [53, 61] to adopt overall accuracy as the evaluation metric.

Implementation Details. We implement CMPAE on the JoMoLD [6] backbone. Following previous approaches [52, 53], we adopt pre-trained VGGish [19] to yield audio features, and employ pre-trained feature extractors, *i.e.* ResNet152 [18] and R(2+1)D [54] for LLP and AVEP and VGG-19 [51] for AVE, to obtain the low-level visual feature representations. The evidence collector f_c^m, $\hat{f}_c^m$ are part of the backbone network with two additional fully-connected layers, activated by LeakyReLU, and the evidence function g is set to Exp. The feature dimension D and the number of snippets are set to 512 and 10. For the stability of training, we select the negative logarithm of the marginal likelihood as the EDL optimization objective in implementation. Specifically, we replace the digamma function in Eq. (5) with a logarithm function. Our model is implemented with Python 3.10 and PyTorch 1.12.1, and we utilize Adam with a batch size of 128 and a learning rate of 5×10^{-4}, which decays by a factor of 0.25 every 6 epochs, for model optimization. We train the model for 25 epochs on all the datasets. Experiments are conducted on a RTX 3090 GPU.

Table 2. AVVP performance comparison with existing methods on the LLP dataset.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVE [53], ECCV2018	49.9	37.3	37.0	41.4	43.6	43.6	32.4	32.6	36.2	37.4
AVSDN [29], ICASSP2019	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [52], ECCV2020	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
CVCMS [30], NeurIPS2021	60.8	63.5	57.0	60.5	59.5	53.8	58.9	49.5	54.0	52.1
MA [58], CVPR2021	59.8	57.5	52.6	56.6	56.6	52.1	54.4	45.8	50.8	49.4
DHHN [22], MM2022	61.4	63.4	56.8	60.5	59.5	54.6	60.8	51.1	55.5	53.3
MM-Pyramid [65], MM2022	61.1	60.3	55.8	59.7	59.1	53.8	56.7	49.4	54.1	51.2
CMBS* [61], CVPR2022	60.2	54.3	50.0	54.8	55.7	51.1	50.8	43.7	48.5	48.3
JoMoLD [6], ECCV2022	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
CMPAE(Ours)	**64.2 (+2.9)**	**66.4 (+2.6)**	**59.2 (+2.0)**	**63.3 (+2.5)**	**62.8 (+2.9)**	**56.6 (+2.7)**	**63.7 (+3.8)**	**51.8 (+2.2)**	**57.4 (+2.9)**	**55.7 (+3.2)**

Table 3. AVE performance comparison.

Methods	Accuracy(%)
AVEL [53], ECCV2018	66.7
AVRB [47], WACV2020	68.9
CMRAN [62], MM2020	72.9
PSP [70], CVPR2021	73.5
CMAN [63], AAAI2022	70.4
MM-Pyramid [65], MM2022	73.2
CMBS [61], CVPR2022	74.2
DPNet [48], ECCV2022	74.5
CMBS [61], fully-supervised	79.3
JoMoLD* [6], ECCV2022	71.8
CMPAE(Ours)	**74.8**

* denotes the reproduced results.

Table 4. AVEP performance comparison with existing methods.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
CMBS [61], CVPR2022	58.0	56.2	52.3	55.5	54.8	51.5	53.6	46.4	50.5	49.4
JoMoLD [6], ECCV2022	60.6	58.9	54.5	58.0	57.7	53.6	55.8	48.6	52.7	51.0
CMPAE(Ours)	**64.1 (+3.5)**	**64.4 (+5.5)**	**58.8 (+4.3)**	**62.4 (+4.4)**	**62.2 (+4.5)**	**57.2 (+3.6)**	**61.9 (+6.1)**	**52.3 (+3.7)**	**57.1 (+4.4)**	**55.6 (+4.6)**

Table 5. Ablation studies of our method.

EDL	PAEC	JML	Seg-level Type		Eve-level Type	
			AVVP	AVEP	AVVP	AVEP
✗	✗	✗	60.8	58.0	54.5	52.7
✓	✗	✗	61.0	58.9	54.9	53.8
✓	✓	✗	61.9	61.5	56.1	55.9
✓	✗	✓	61.4	60.8	55.3	54.6
✓	✓	✓	**63.3**	**62.4**	**57.4**	**57.1**

4.2. Comparison with State-of-the-art Methods

Evaluation on AVVP. As shown in Table 2, CMPAE outperforms previous weakly supervised methods in all metrics on the AVVP dataset. Compared with our backbone JoMoLD [6], the absolute gains are also remarked. Specifically, our approach achieves favorable performance of 63.7% and obtains absolute gains of 9.3% and 3.8% in terms of the event-level visual F-score when compared to the SOTA approaches MA [58] and JoMoLD. Note that some other methods, such as CVCMS and DHHN [22, 30] also consider joint single-modal and cross-modal learning, however, they neglect the relation constraint between single-modality and joint-modality for explicit cooperation and calibration learning. As a result, our method outperforms them by a considerable margin.

Evaluation on AVE. CMPAE is compared against several SOTA methods in Table 3, showing superior or comparable performance. Notably, our approach outperforms the representative methods PSP [70] and CMBS [61] by 1.3% and 0.6%. For the recent work DPNet [48], we also achieve a comparable performance. From the table, we also observe that the performance of the weakly-supervised methods is close to that of the fully-supervised one, showing that there is a bottleneck for further improvement on this task.

Evaluation on AVEP. After validating on AVVP and AVE separately, we then conduct experiments on the combined dataset, which is referred to as AVEP. In this dataset, the metrics of the AVVP task are adopted due to the evaluation diversity. Although combing two datasets is not a contribution, performing evaluations on two datasets with specific properties jointly can help us investigate the capability and generalization ability of the compared approaches. Here, we employ the two representative and SOTA baselines from AVVP and AVE, including CMBS and JoMoLD, for comparison. From Table 4 we can observe that our proposed CMPAE approach achieves superior performance. Note that compared with the dataset of AVVP, although 14 new categories are added, the performance of our method is still comparable, showing favorable model capability.

4.3. Further Remarks

Effectiveness of the Presence-Absence Evidence Collector (PAEC) and Joint-modal Mutual Learning (JML). PAEC and JML are the two essential components of our method. To verify the effectiveness, we first design a fundamental baseline, which equips the backbone network with only the vanilla EDL objective [50]. Then we progressively add PAEC and JML to the baseline, as shown in Table 5, the corresponding performance consistently increases, proving the positive impact of the presence-absence evidence collecting and mutual learning between uni-modality and joint-modality. In addition, PAEC and JML can enhance and cooperate with each other for a more significant performance.

In-depth Analysis of PAEC. In PAEC, we utilize the features of a single modality to learn presence evidence, while the absence evidence is collected with the help of complementary modality features. As shown in Table 6, only using a single track for generating the corresponding presence-absence evidence is insufficient, which validates the effectiveness of considering complementary information. In addition to this analysis, there are still another two questions:

(1) *Why not leverage cross-modal information for both presence and absence evidence learning?* Similar to absence evidence learning, complementary cross-modal information is also useful for the presence evidence learning in the current modality. However, as shown in Table 6, if we indiscriminately use cross-modal features for the prediction of a single track, the modality-specific information is prone to being overlooked, which hinders the perception of individual audio or visual events. Moreover, the mutual learning will be disturbed due to the information homogeneity.

(2) *Why not exchange the feature sources for learning presence and absence evidence?* As stated above, the cross-modal information is also beneficial for presence evidence learning. However, to improve the discriminative ability of a single track, it is desirable that the modality itself can provide sufficient presence evidence for target categories, which can be regarded as a regularization for model learning. While for absence learning, the other modality should provide reference signals but not dominant ones, which are partially admitted by [66]. Therefore, our strategy for presence-absence learning is reasonable, as it can not only mine the potential of a single modality but also benefit from the complementary track on a limited scale. Table 6 shows that the exchanged setting makes the performance decrease.

In-depth Analysis of JML. Different from the traditional mutual learning [67], in our proposed method, uncertainty of evidence is adopted for more reliable learning. To explore its effectiveness, as shown in Table 6, we abandon the uni-modal uncertainty term (denoted as w/o u^{uni}) or the joint-modal uncertainty term (denoted as w/o u^{av}) in our framework, which shows degraded performance. In addition, we adopt the Mean operator uniformly for the target and non-target categories (denoted as w/o $\delta(c)$) to validate the effectiveness of the uni-modal selector $\delta(c)$. The results show that our proposed JML can effectively leverage the relations between single-modality and joint-modality.

Qualitative Analysis. To further analyze the reliability of the learned presence-absence evidence, the qualitative results are shown in Figure 3. For the audio and visual tracks of each video, we highlight the corresponding presence and absence evidence values of each snippet. We can observe that in most cases, the presence evidence of the target class is high while the absence evidence is low, and vice versa. Moreover, the evidence from different modalities has a clear distinction when only a single-modality event occurs.

Table 6. In-depth analysis of our proposed PAEC and JML.

Models	Seg-level Type		Eve-level Type	
	AVVP	AVEP	AVVP	AVEP
both uni-modal	61.2	60.9	55.2	54.4
both cross-modal	61.7	61.3	55.7	55.9
exchange uni/cross	62.1	61.6	56.4	56.0
w/o u^{av}	62.2	61.8	56.5	56.3
w/o u^{uni}	62.0	61.7	56.4	56.0
w/o $\delta(c)$	62.5	61.8	56.3	56.5
CMPAE	**63.3**	**62.4**	**57.4**	**57.1**

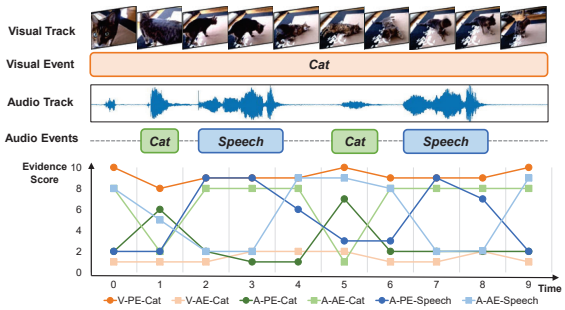

Figure 3. The presence-absence evidence strength of a test video that contains event categories of *Cat* and *Speech*. 'V', 'A', 'PE', and 'AE' denote the visual modality, audio modality, presence evidence, and absence evidence, respectively. Note that the numbers of evidence strength are rounded for better illustration.

5. Conclusions

This paper proposes a cross-modal presence-absence evidence learning method for WS-AVEP, which jointly enjoys the merits of uni-modal discrimination and cross-modal enhancement under Subject Logic theory. Specifically, the cooperative presence-absence evidence collector and the joint-modal mutual learning module are capable of learning and calibrating reliable evidence. The encouraging performance is validated in extensive experiments. In this work, although we attempt to combine two datasets together for a more comprehensive experiment, larger scale datasets with considerable diversity are still expected to be used in the future to evaluate the effectiveness of different approaches.

Acknowledgements

This work was supported by the National Key Research & Development Plan of China under Grant 2020AAA0106200, in part by the National Natural Science Foundation of China under Grants 62036012, 62236008, U21B2044, 62102415, 61721004, 62072286, 62072455, 62002355, in part by Beijing Natural Science Foundation (L201001), and in part by Open Research Projects of Zhejiang Lab (NO.2022RC0AB02).

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, 2020. 2, 3

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 2

[3] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV (ICCV)*, 2021. 2, 3, 4

[4] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016.

[5] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *ECCV*, 2022. 2, 3, 4

[6] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. *ECCV*, 2022. 2, 3, 4, 5, 6, 7

[7] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19999–20009, June 2022. 3

[8] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 2

[9] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019. 2

[10] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 2

[11] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Smart: Joint sampling and regression for visual tracking. *IEEE Transactions on Image Processing*, 28(8):3923–3935, 2019. 1

[12] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3476–3491, 2021. 2

[13] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 2

[14] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 2

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 6

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*. PMLR, 2017. 3

[17] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *ICLR*, 2020. 2, 3, 4

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 6

[20] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019. 2

[21] Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. Multidimensional uncertainty-aware evidential neural networks. In *AAAI*, 2021. 3

[22] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *ACM MM*, 2022. 2, 3, 7

[23] Audun Jsang. Subjective logic: A formalism for reasoning under uncertainty. *Springer Verlag*, 2016. 2, 3, 4, 5

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019. 2

[25] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *ICCV*, 2019. 2

[26] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*, 2020. 2

[27] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, 2022. 2, 3, 4

[28] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 3

[29] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*, 2019. 3, 7

[30] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *NeurIPS*, 2021. 1, 2, 3, 6, 7

[31] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *ACCV*, 2020. 3

[32] Shuo Liu, Weize Quan, Yuan Liu, and Dong-Ming Yan. Bi-directional modality fusion network for audio-visual event localization. In *ICASSP*, 2022. 3

[33] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted deep learning with opinion aggregation. In *AAAI*, 2022. 3

[34] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *ICLR*, 2020. 2

[35] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. *arXiv preprint arXiv:2210.05060*, 2022. 3

[36] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018. 2, 3, 4

[37] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, 2022. 3

[38] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Advances in Neural Information Processing Systems*, 2022. 3

[39] Shentong Mo and Yapeng Tian. Semantic-aware multi-modal grouping for weakly-supervised audio-visual video parsing. In *ECCV Workshop*, 2022. 3

[40] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *CVPR*, 2021. 2

[41] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 2

[42] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 3

[43] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. In *NeurIPS*, 2019. 2

[44] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2

[45] Deep Shankar Pandey and Qi Yu. Multidimensional belief quantification for label-efficient meta-learning. In *CVPR*, 2022. 3

[46] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP*. IEEE, 2020. 3

[47] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *WACV*, 2020. 1, 3, 7

[48] Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Dual perspective network for audio-visual event localization. In *ECCV*, 2022. 3, 7

[49] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 2

[50] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018. 2, 3, 4, 7

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[52] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7

[53] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 2, 3, 6, 7

[54] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 6

[55] Dennis Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021. 2, 3

[56] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 1

[57] Yu Kong Wentao Bao, Qi Yu. Opental: Towards open set temporal action localization. In *CVPR*, 2022. 2, 3

[58] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021. 2, 3, 4, 7

[59] Yiling Wu, Xinfeng Zhang, Yaowei Wang, and Qingming Huang. Span-based audio-visual localization. In *ACM MM*, 2022. 3

[60] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019. 1, 2, 3

[61] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *CVPR*, 2022. 1, 2, 3, 6, 7

[62] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *ACM MM*, 2020. 1, 3, 7

[63] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*, 2020. 1, 3, 7

[64] Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008. 2, 3

[65] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM MM*, 2022. 3, 7

[66] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. Audio-adaptive activity recognition across video domains. In *CVPR*, 2022. 1, 2, 3, 8

[67] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 8

[68] Xujiang Zhao, Xuchao Zhang, Wei Cheng, Wenchao Yu, Yuncong Chen, Haifeng Chen, and Feng Chen. Seed: Sound event early detection via evidential uncertainty. In *ICASSP*, 2022. 4

[69] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019. 2

[70] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *CVPR*, 2021. 3, 7

[71] Yonggang Zhu, Chao Tian, Zhuqing Jiang, Aidong Men, Haiying Wang, and Qingchao Chen. Mixed in time and modality: Curse or blessing*f* cross-instance data augmentation for weakly supervised multimodal temporal fusion. In *ICASSP*, 2022. 3