Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing

Haoyue Cheng[1,2]   Zhaoyang Liu[2]   Hang Zhou[3]   Chen Qian[2]
Wayne Wu[2]   Limin Wang[1,4 ✉]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] SenseTime Research   [3] CUHK - Sensetime Joint Lab   [4] Shanghai AI Laboratory
chenghaoyue98@gmail.com   zyliumy@gmail.com   zhouhang@link.cuhk.edu.hk
qianchen@sensetime.com   wuwenyan0503@gmail.com   lmwang@nju.edu.cn

Abstract. This paper focuses on the weakly-supervised audio-visual video parsing task, which aims to recognize all events belonging to each modality and localize their temporal boundaries. This task is challenging because only overall labels indicating the video events are provided for training. However, an event might be labeled but not appear in one of the modalities, which results in a modality-specific noisy label problem. In this work, we propose a training strategy to identify and remove modality-specific noisy labels dynamically. It is motivated by two key observations: 1) networks tend to learn clean samples first; and 2) a labeled event would appear in at least one modality. Specifically, we sort the losses of all instances within a mini-batch individually in each modality, and then select noisy samples according to the relationships between intra-modal and inter-modal losses. Besides, we also propose a simple but valid noise ratio estimation method by calculating the proportion of instances whose confidence is below a preset threshold. Our method makes large improvements over the previous state of the arts (*e.g.*, from 60.0% to 63.8% in segment-level visual metric), which demonstrates the effectiveness of our approach. Code and trained models are publicly available at https://github.com/MCG-NJU/JoMoLD.

Keywords: Audio-visual video parsing; multi-modal learning; weakly-supervised learning; label denoising

1   Introduction

Many works show that the audio-visual clues play a crucial role in comprehensive video understanding [3,45,30]. However, most studies on audio-visual joint learning [38,17,31] assume that the two modalities are correlated or even temporally synchronized, which is not always the case. For example, the sound of a
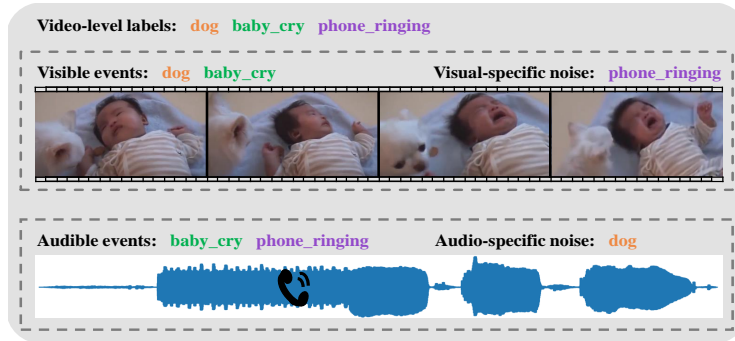

✉: Corresponding author.

Fig. 1: **An example to illustrate modality-specific noise in weakly-supervised AVVP task.** An infant accompanied by a dog is sleeping. Then, the phone bell rings and it frightens the baby to cry. In this example, the visible events in the whole video are "dog" and "baby_cry_infant_cry". The audible events are "telephone_bell_ringing" and "baby_cry_infant_cry". Thus, the "dog" can be treated as audio-specific noise, and "telephone_bell_ringing" is regarded as visual-specific noise.

car might be out of sight, but such information is still crucial for real-world perception. To this end, Tian *et al.* [37] proposed the **audio-visual video parsing** (AVVP) task without consistency restriction, which aims to recognize all events belonging to each modality and localize their temporal boundaries.

AVVP task is formulated in a weakly-supervised manner since precisely annotating labels would be expensive and time-consuming. Event labels are provided for each video in the training set, but the events' detailed modality and temporal location are unavailable. This manner is termed Multimodal Multiple Instance Learning (MMIL). The weakly-supervised setting and audio-visual inconsistency lead to a serious issue: **modality-specific noise**. Modality-specific noise is related to *the event clues that do not appear in one of the two modalities*. Taking Figure 1 as an example, the event "telephone_bell_ringing" only appears in the audio track and can not be perceived from the visual track, which thus is an improper supervised signal for visual modality and can be regarded as a visual-specific noisy label. We argue that lowering the interference of modality-specific noise can significantly advance the quality of audio-visual video parsing.

The pioneers in audio-visual video parsing have made sustained efforts to learn the spatio-temporal clues under such a weakly-supervised setting. HAN [37] adopts a hybrid attention network optimized with a label smoothing mechanism, but it still suffers from the modality-specific noise. MA [43] exchanges audio or visual tracks between two unrelated videos to yield reliable event labels for each modality. However, since the cross-modal aggregation in [43] is trained by paired audio-visual features, this might interfere with the uncertainty-assessing procedure on their re-assembled videos. As the previous label denoising methods [42,47,29,21] focus on unimodal patterns, leveraging the cross-modal cor-

relation to alleviate modality-specific noise still remains to be studied further. Consequently, we design a novel training strategy to mitigate the impacts of modality-specific noise for weakly-supervised audio-visual video parsing.

In this paper, we propose the **Joint-Modal Label Denoising (JoMoLD)** training strategy to dynamically alleviate **modality-specific noise** through careful loss analysis for both modalities. We take the inspiration from two observations. 1) Neural networks tend to learn cleanly labeled samples first, and gradually memorize noisy ones [48,6,22]. As a result, most noisily labeled data would be more challenging to learn than correctly labeled ones. From the view of loss patterns, the loss of cleanly labeled samples would be lower than noisily labeled ones. 2) Under the weakly supervised training setting, an event label ought not to serve as noise for both modalities, *i.e.*, this event appears in at least one modality. According to these observations, the loss of a noisily labeled modality tends to be higher than the loss of the other modality where the event appears. We call this phenomenon *loss inconsistency among different modalities*.

Based on the above analysis, we leverage audio and visual loss patterns to remove modality-specific noisy labels for each modality. First, we design a noise estimator to approximately pre-estimate the noise ratio per category individually for each modality, which guides our proposed training strategy in determining the modality-specific noise. Then, when training the parsing model, we rank the losses within a mini-batch separately for two modalities. Based on the pre-estimated noise ratios, we treat the labels with inconsistent losses between two modalities as modality-specific noise. For each iteration in the training phase, we remove modality-specific noisy labels from their corresponding modality before back-propagation. This training strategy makes the parsing network robust to modality-specific noise under the weakly-supervised setting of AVVP. Our experiment results significantly outperform the previous state of the arts, which validates the effectiveness of our proposed method.

In summary, we make the following contributions:

- *First*, we develop a noise ratio estimator to calculate the modality-specific noise ratios, which play a crucial role in determining the noise selection.
- *Second*, we propose a general and dynamic training paradigm, namely Joint-Modal Label Denoising (JoMoLD), to alleviate the issue of modality-specific noise from the perspective of joint-modal label denoising on weakly-supervised AVVP task.
- *Finally*, the experiments on the LLP dataset validate the effectiveness of our method. Especially, the segment-level visual metric is improved from 60.0% to 63.8% over the state of the art.

2   Related Work

2.1   Audio-Visual Learning

Audio-visual joint learning has derived a variety of tasks, such as audio-visual representation learning [7,12,3,20,1], audio-visual sound separation [10,9,50,51],

sound source localization [33,4], audio-visual video captioning [32,16,36], audio-visual event localization [23,44,52], and audio-visual action recognition [45,30,11]. Most of these works are based on the assumption that audio and visual signals are always semantically corresponding and temporally synchronized.

For audio-visual representation learning, [7] and [12] transfer discriminative visual knowledge from pre-trained visual models into the audio modality. Alwassel *et al.* [1] leverage unsupervised clustering results within one modality as a supervised signal to the other modality. Other works explore learning instance-level audio-visual correspondence. Arandjelovic *et al.* [3] design a pretext task to learn correspondent representations between images and audio. Korbar *et al.* [20] further utilize the temporal synchronization between audio and visual streams.

2.2   Weakly-Supervised AVVP

Audio-visual video parsing (AVVP) task [37] breaks the restriction that audio and visual signals are definitely aligned. Due to the difficulty of exhaustive manual annotations, the audio-visual video parsing task is under the weakly-supervised setting, where only the video-level labels are provided for training.

To tackle the weakly-supervised AVVP task, previous work [37] proposes a hybrid attention network and attentive Multimodal Multiple Instance Learning (MMIL) Pooling mechanism to aggregate all features. Wu *et al.* [43] refine individual modality labels by swapping audio or visual tracks between two unrelated videos. Nevertheless, they independently refine modality labels without considering the relationship between the two modality labels. Our work obtains more precise modality-specific labels in a joint-modal label denoising manner.

2.3   Learning With Label Noise

Deep neural networks have been demonstrated to learn clean samples first, and gradually memorize samples with noisy labels [48,6,22,13]. Recent works [25] show that such early-learning and memorization phenomena can even be observed in linear models. Many works are trying to handle this problem from different perspectives. One kind of methods [49,40,19] designs reasonable loss functions or regularization mechanisms to reduce overfitting noise. Semi-supervised learning is also adopted in some works [21,27].

Two types of methods related to our approach are learning on selected clean samples [26,13,47,29] and correcting noisy labels [5,35,46,34]. Han *et al.* [13] propose "Co-teaching", which optimizes two models on clean samples selected by the paired network. To prevent the two networks converging to a consensus, "Co-teaching+" [47] is proposed to combine "Co-teaching" with "Update by Disagreement" [26] strategy. These collaborative learning with label noise methods are motivated by "Co-training" [8] in semi-supervised training. Tanaka *et al.* [35] utilize model output to reassign labels for noisy samples. However, most of these works focus on single-modal label denoising, and cannot be directly utilized for multi-modal label denoising.

Some multi-modal learning works [2,18] also try to mitigate the effects of noisy labels. Amrani *et al.* [2] propose to reduce multi-modal noise estimation to a multi-modal density estimation problem. Hu *et al.* [18] propose a "Robust Clustering loss" to make the model focus on clean samples instead of noisy ones. However, these methods do not explore the correlation between multi-modalities and have not addressed the multi-modal noisy labels problem.

Our work deals with the issue of modality-specific noisy labels in weakly-supervised AVVP task. We aim to generate more reliable modality-specific labels considering cross-modal connections.

3   Method

Our main idea is to utilize the cross-modal loss patterns to perceive modality-specific noise, which is compatible with off-the-shelf networks (*e.g.*, [37]) on weakly-supervised AVVP task. In this section, we elaborately introduce our proposed method, *i.e.*, **Joint-Modal Label Denoising** (JoMoLD). Firstly, we formulate the problem statement along with the baseline framework in Sec. 3.1. Secondly, the noise estimator is proposed to calculate modality-specific noise ratios in Sec. 3.2. Thirdly, we design an effective algorithm to remove modality-specific noisy labels by pre-estimated noise ratios in Sec. 3.3. Lastly, we give an insightful discussion of our method in Sec. 3.4.

3.1   Preliminaries

Problem Statement. This task aims to detect audible events or visible events that appear in each segment of a video, which is formulated as a Multimodal Multiple Instance Learning (MMIL) problem with C event categories by Tian *et al.* [37]. Specifically, given a T-second video sequence $\{A_t, V_t\}_{t=1}^{T}$, A_t denotes t-th segment in audio track and V_t denotes t-th segment in visual track. During evaluation, let (y_t^a, y_t^v, y_t^{av}) denote audio, visual and audio-visual event labels at t-th segment, respectively. Note that y_t^a, y_t^v and y_t^{av} are $\{0,1\}^C$ vectors, indicating the presence or absence of each event category. The audio-visual event represents this event occurs in both visual track and audio track at time t. However, as a weakly-supervised task, we can only access video-level event label $y \in \{0,1\}^C$ instead of accurate segment-level labels during training. In other words, we only know which events occurred in a video, but can not acquire when events occur and in which modality the events appear. Following the practice in [37], we use pre-trained off-the-shelf networks to extract local audio and visual features $\{f_t^a, f_t^v\}_{t=1}^{T}$ for each segment.

Baseline Framework. We here use previous work [37] (denoted as $\mathcal{F}$) as our important baseline. To capture temporal context and leverage the clues in different modalities, Tian *et al.* [37] adopt self-attention and cross-attention mechanisms to aggregate inner-modal and intra-modal information on segment features. Furthermore, an attentive MMIL Pooling mechanism is proposed to yield modality-level and video-level predictions. As a multi-label multi-class learning

task, it naturally uses binary cross-entropy (BCE) loss to optimize the model. In the baseline, the video-level label y is used to supervise both modality-level and video-level predictions. Due to the modality-specific noise, we argue that such a fashion is misleading for model training. Therefore, we develop a dynamic training strategy, termed **Joint-Modal Label Denoising** (JoMoLD), to alleviate the effect of modality-specific noise during training.

3.2   Estimating Noise Ratios

Our algorithm requires pre-estimating the modality-specific noise ratio per category, which assists our model in determining which labels should be removed during training. Since the real modality-specific noise ratios are unavailable in the training phase, we design a simple yet effective manner to approximately estimate noise ratios. The key insight is that the baseline model trained on this task already has a certain capacity for discriminating noisy labels.

Specifically, we train a noise estimator $\mathcal{H}$ following baseline [37] with one notable modification: removing the cross-modal attention in the overall pipeline. As cross-modal attention exchanges the information between audio and visual features, it practically interferes with the noise estimation for each modality. Experiment in Table 2c has proved this point. Let $\hat{\mathbf{P}}^a, \hat{\mathbf{P}}^v \in \mathbb{R}^{N \times C}$ denote the audio and visual predictions of $\mathcal{H}$, and $\bar{P}^a, \bar{P}^v \in \mathbb{R}^C$ are the mean of predictions for each category. $\mathbf{Y} \in \{0,1\}^{N \times C}$ are video-level labels. Note that N is the number of videos in training set. Then, we need to define which labels are probably noise. For example, we argue the annotated c-th category label for i-th video, i.e. $\mathbf{Y}[i,c] = 1$, is not reliable for audio modality if $\hat{\mathbf{P}}^a[i,c]/\bar{P}^a[c]$ is lower than a pre-set threshold θ^a. Here, $\hat{\mathbf{P}}^a[i,c]$ is normalized by $\bar{P}^a[c]$ so as to alleviate the impact of imbalanced distribution of predictions in each event category. The procedure of estimating noise ratio for c-th category is summarized as follows:

$$
\begin{aligned}
\mathbf{r}^a[c] &= \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{\mathbf{P}}^a[i,c]/\bar{P}^a[c] < \theta^a) \times \mathbf{Y}[i,c]}{\sum_{i=1}^{N} \mathbf{Y}[i,c]}, \\
\mathbf{r}^v[c] &= \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{\mathbf{P}}^v[i,c]/\bar{P}^v[c] < \theta^v) \times \mathbf{Y}[i,c]}{\sum_{i=1}^{N} \mathbf{Y}[i,c]},
\end{aligned}
\tag{1}
$$

where $\mathbf{r}^a$ ($\mathbf{r}^v$) $\in \mathbb{R}^C$ denotes noise ratios of positive labels in audio (visual) track for C categories, $\mathbb{I}$ is the indicator function, and "$\times$" denotes multiplication. Note that θ^a (θ^v) can be seen as the confidence to determine noisy labels for audio (visual) modality. Table 2b shows that final performance of our proposed method is not sensitive to θ^a and θ^v. The estimated noise ratios will be used as a priori knowledge for the label denoising procedure.

Generally, our noise estimator is essentially different from Wu *et al.* [43]. They exchange the audio tracks between two unrelated videos to filter out the noisy labels and re-train the model from scratch based on refined labels. On the one hand, as cross-modal attention aggregates multi-modal clues, the modality-level predictions would affect each other. Thus it is improper to assess the uncertainty
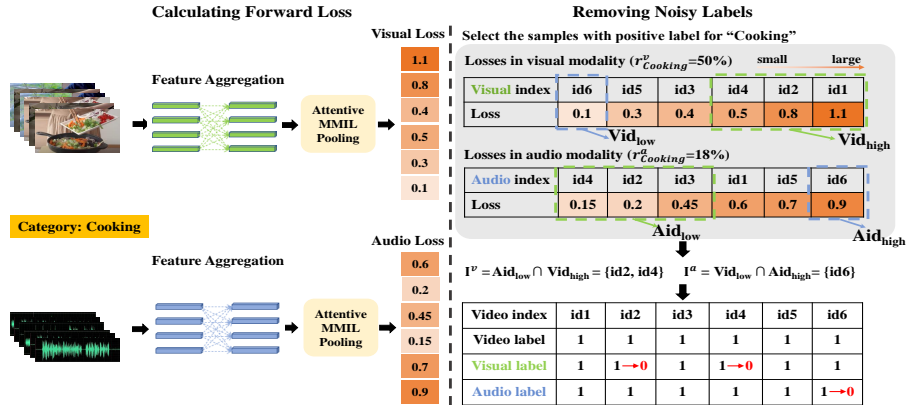
Fig. 2: **The proposed modality-specific label denoising procedure.** The label denoising procedure consists of **Calculating Foward Loss** and **Removing Noisy Labels**. In this case, we represent the label denoising process for the "Cooking" event in a batch of videos. In calculating forward loss, we aggregate the intra-modal features, obtain the modality-level predictions, and further calculate the modality losses. Based on the estimated noise ratios $\mathbf{r}^v_{Cooking}$ and the sorted visual losses, we obtain the indices of noisy samples for visual modality, *i.e.* $\mathbf{I}^v$. Then we remove the video labels for videos by $\mathbf{I}^v$ to generate refined visual labels. In the figure, "$\rightarrow 0$" denotes the label of "Cooking" is removed. The same procedure is applied to audio modality label denoising.

of event labels for each modality when using cross-modal attention. On the other hand, the refined labels will be fixed after the label refinement procedure in [43]. It causes that even the wrongly refined labels would also be used to re-train the parsing model in the whole re-training phase. In contrast, our method alleviates the potential negative impacts of cross-modal attention in the phase of noise estimation. Furthermore, modality-specific noisy labels are removed dynamically, which is expected to tolerate some improper refinements in the previous training. Even though the intrinsic biases might exist in the noise estimator, the estimated noise ratios are still effective for our modality-specific label denoising algorithm, which is also validated in experiments.

3.3   Modality-specific Label Denoising

In this section, our goal is to remove modality-specific noisy labels and optimize the parsing model $\mathcal{F}$ with the refined labels. Therefore, we need to solve a tricky issue: how to identify the modality-specific noisy labels. As illustrated in Figure 2, our modality-specific label denoising procedure can be summarized as two steps: a) **Calculating Forward Loss** and b) **Removing Noisy Labels**. As a

---

Algorithm 1 The Pipeline of JoMoLD

Require:

    Noise ratios $\mathbf{r}^a \in \mathbb{R}^C, \mathbf{r}^v \in \mathbb{R}^C$ estimated in Sec. 3.2

    Total training iterations Γ, the number of categories C, the batch size B

    The parsing network $\mathcal{F}$

1: **for** $i = 0$ to Γ - 1 **do**

2:     Fetch a mini-batch $\mathcal{B}$, and video-level labels $\mathbf{Y} \in \{0,1\}^{B \times C}$, $\mathbf{Y}^a = \mathbf{Y}$, $\mathbf{Y}^v = \mathbf{Y}$

3:     Feed $\mathcal{B}$ into $\mathcal{F}$ (skipping cross-modal attention) to calculate forward loss $\mathbf{L}^a$, $\mathbf{L}^v \in \mathbb{R}^{B \times C}$

4:     **for** $c = 1$ to C **do**

5:         # Find the indexes of positive labels and the number of positive samples

            $\mathbf{I} = nonzero(\mathbf{Y}[:,c]), \quad B' = \sum_{i=1}^{B} \mathbf{Y}[:,c]$

        # Calculate the numbers of candidate noise for audio and visual modalities

            $\mathbf{M}^a = int(\mathbf{r}^a[c] \times B'), \quad \mathbf{M}^v = int(\mathbf{r}^v[c] \times B')$

        # Determine the indexes of audio noise in batch $\mathcal{B}$

            $\mathbf{I}^a = \mathbf{I}[\mathcal{G}(-\mathbf{L}^a[\mathbf{I},c], \mathbf{M}^a)] \cap \mathbf{I}[\mathcal{G}(\mathbf{L}^v[\mathbf{I},c], \mathbf{M}^a)]$

        # Determine the indexes of visual noise in batch $\mathcal{B}$

            $\mathbf{I}^v = \mathbf{I}[\mathcal{G}(-\mathbf{L}^v[\mathbf{I},c], \mathbf{M}^v)] \cap \mathbf{I}[\mathcal{G}(\mathbf{L}^a[\mathbf{I},c], \mathbf{M}^v)]$

6:         # Remove noisy labels

            $\mathbf{Y}^a[\mathbf{I}^a,c] = 0, \quad \mathbf{Y}^v[\mathbf{I}^v,c] = 0$

7:     **end for**

8:     Feed $\mathcal{B}$ into $\mathcal{F}$, and utilize $\mathbf{Y}$, $\mathbf{Y}^a$ and $\mathbf{Y}^v$ to optimize network $\mathcal{F}$

9: **end for**

---

general training strategy, we adopt the model in [37] as our backbone network $\mathcal{F}$ to verify our method.

First, in the step of **Calculating Forward Loss**, losses are calculated in each modality for removing the noisy labels. Specifically, we feed the extracted local features of a batch of videos into the network, and calculate the BCE losses for each modality. Here, cross-modal attention is skipped in this step to avoid interference of cross-modal feature aggregation. Note that the losses calculated in this step are only utilized to remove noisy labels but not optimize the network. Let B represent the batch size, $\mathbf{L}^a, \mathbf{L}^v \in \mathbb{R}^{B \times C}$ denote the BCE losses in audio and visual modalities.

Second, in the step of **Removing Noisy Labels**, modality-specific noisy labels are determined according to the loss patterns based on estimated noise ratios $\mathbf{r}^a$ and $\mathbf{r}^v$. We here define a function for better introducing the procedure of removing labels:

$$\mathcal{G}(\mathcal{L}, n) = argsort(\mathcal{L})[0:n], \tag{2}$$

where $\mathcal{L}$ denotes the losses of a batch of samples, and n is the parameter to control the number of selected indexes. $argsort(\mathcal{L})$ is a function that sorts $\mathcal{L}$ in ascending order and returns the indexes of sorted losses. We use $argsort(-\mathcal{L})$ to obtain the indexes of losses $\mathcal{L}$ sorted in descending order. The $nonzero(\cdot)$ is a function that returns the indexes of samples whose value is not 0. The procedure to determine which labels are modality-specific noise corresponds to Step 5 in

Algorithm 1. Intuitively, taking audio-specific label denoising as an example, we argue that a positive label with a high loss in the audio modality and a low loss in visual modality would be an audio-specific noisy label. We have discussed the interpretability and reasonability of the way to determine noisy labels in Sec. 3.4. The identified modality-specific noisy labels are removed from $\mathbf{Y}$, to generate refined modality labels $\mathbf{Y}^a$ and $\mathbf{Y}^v$.

Lastly, in each training iteration, we feed the batch $\mathcal{B}$ to network $\mathcal{F}$ to obtain predictions. The refined labels $\mathbf{Y}^a$, $\mathbf{Y}^v$ and $\mathbf{Y}$ serve as supervised signals for audio predictions, visual predictions, and video predictions, respectively. Experimental results in Sec. 4 have validated the effectiveness of our method.

3.4   Discussion

As we have mentioned above, we first train the noise estimator $\mathcal{H}$ to estimate the modality-specific noise ratios for each category. $\mathcal{H}$ is modified from the baseline $\mathcal{F}$ [37] by removing cross-modal attention. After that, we adopt the noise ratios to guide the training process. We use the network $\mathcal{F}$ as our parsing model, but skip the cross-modal attention during **calculating forward loss**. Finally, there is still confronted with two critical questions about the motivation of our method:

1) Why do we regard the event labels with higher losses as the candidate noise set by using intra-modal loss patterns? As analyzed previously, deep neural networks are prone to learn clean labels first, but over-fit the noisy labels with more training epochs [6,48,22]. In other words, the losses of clean labels would drop faster than noisy labels. Built upon this observation, we argue that noisy labels are more likely to exist in the samples with higher losses.

2) Why do cross-modal cues make more evident improvement for determining the modality-specific noise? If we directly treat the event labels with higher losses as noise, some hard labels would not be seen by the model during training in extreme circumstances. We observe that a video-level event label typically appears in at least one modality. In this sense, if one label has a higher loss in audio modality but a lower loss in the visual modality, it means evident clues have appeared in visual rather than audio modality. Therefore, we speculate that this label may be noisy for audio modality with high confidence. It is effective to utilize the complementary knowledge to recheck label noise, which is also verified in Table 2d.

4   Experiments

This section elaborates on our experiments' details and compares our proposed JoMoLD with state-of-the-art methods. In the ablation studies, we present the effect of each module. We conduct a qualitative analysis and show the advantages of our JoMoLD over state-of-the-art methods.

4.1   Experiment Settings

Dataset. We evaluate our method for weakly-supervised AVVP task on the *Look, Listen, and Parse* (LLP) dataset. It consists of 11849 10-second videos with 25 event categories. The categories cover a wide range of domains such as human activities, animal activities, music performance, *etc.* We utilize the official data split for training and evaluation. There are 10000 videos for training and 1849 validation-test videos for evaluation.

Evaluation Metrics. We evaluate the parsing performance of all events (audio, visual, and audio-visual events) under segment-level and event-level metrics. We use both segment-level and event-level F-scores as metrics. The former metrics evaluate the segment-wise prediction performance. The latter metrics are designed to extract events with consecutive positive snippets in the same event categories, and 0.5 is used as the mIoU threshold to compute event-level F-scores. Moreover, we also assess the comprehensive performance for all events by "Type@AV" and "Event@AV" metrics. Type@AV is calculated by averaging audio, visual, and audio-visual metrics. Event@AV computes the results considering all audio and visual events instead of averaging metrics. Abbreviations of metric names are represented in all experiment tables, where "A" denotes audio events, "V" represents visual events, "AV" denotes audio-visual events, "Type" indicates Type@AV, and "Event" denotes "Event@AV".

Implementation Details. Following the data preprocessing in [37,43], we decode a 10-second video into 10 segments, and each segment contains 8 frames. We use pre-trained ResNet152 [14] and R(2+1)D [39] to capture the appearance and motion features and concatenate them as low-level visual features. For audio, we adopt pre-trained VGGish [15] to yield audio features. Adam optimizer is used to train the model, and the learning rate 5e-4 drops by a factor of 0.25 for every 6 epochs. We train the model for 25 epochs with batch size 128.

4.2   Comparison with State-of-the-art Methods

We compare our method with different types of methods: weakly-supervised sound event detection methods TALNet [41], weakly-supervised action localization methods STPN [28] and CMCS [24], modified audio-visual event localization methods AVE [38] and AVSD [23], the state-of-the-art AVVP methods HAN [37] and MA [43]. In addition, Co-teaching+ [47] and JoCoR [42] are famous for learning with label noise methods that focus on single-modal tasks but not multi-modal tasks. On this weakly-supervised AVVP task, we reproduce the variants of these two methods [47,42] utilizing the backbone in HAN to compare with our method. The variants of these two methods are denoted as " HAN *w/* Co-teaching+" (*abbr.* HC) and " HAN *w/* JoCoR" (*abbr.* HJ).

Table 1 shows the results of JoMoLD and other state-of-the-art methods on the LLP test dataset. Our JoMoLD here adopts the optimal settings studied by Sec. 4.3. As a label denoising strategy, JoMoLD can combine with other feature learning methods to achieve higher performance, such as contrastive

Table 1: **Comparisons with the state-of-the-art methods on the test set of LLP.** JoMoLD achieves the best performance among them. "CL" denotes the contrastive learning proposed in MA [43]. We simply add "CL" loss into the existing loss functions when optimizing the network, but do not utilize it in label denoising. Results of our method combined with "CL" proves the flexibility and effectiveness of JoMoLD. "-" denotes this result is not available.

Methods	Segment-level					Event-Level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
TALNet [41]	50.0	-	-	-	-	41.7	-	-	-	-
STPN [28]	-	46.5	-	-	-	-	41.5	-	-	-
CMCS [24]	-	48.1	-	-	-	-	45.1	-	-	-
AVE [38]	49.9	37.3	37.0	41.4	43.6	43.6	32.4	32.6	36.2	37.4
AVSDN [23]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [37]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
HAN w/ Co-teaching+ (HC) [47]	59.4	56.7	52.0	56.0	56.3	50.7	53.9	46.6	50.4	48.7
HAN w/ JoCoR (HJ) [42]	61.0	58.2	53.1	57.4	57.7	52.8	54.7	46.7	51.4	50.3
MA [43] w/o CL	59.8	57.5	52.6	56.6	56.6	52.1	54.4	45.8	50.8	49.4
MA [43] w/ CL	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
JoMoLD (Ours) w/o CL	60.6	62.2	56.0	59.6	58.6	53.1	58.9	49.4	53.8	51.4
JoMoLD (Ours) w/ CL	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**	**53.9**	**59.9**	**49.6**	**54.5**	**52.5**

learning proposed in MA [43]. Notably, our method outperforms the state-of-the-art methods (*e.g.*, HC, HJ, and MA) by a non-negligible margin. For example, JoMoLD is higher than MA by 3.8 points in the segment-level visual event parsing metric. These results demonstrate the effectiveness of our strategy of joint-modal label denoising.

4.3 Ablation Studies

This section performs ablation studies on estimating noise ratios and modality-specific label denoising, respectively. Segment-level metrics are reported if not stated. The optimal settings are explored by the following ablations.

Ablation Studies on Estimating Noise Ratios:

Study the Effectiveness of Noise Estimator. To verify the importance of the noise estimator, we compare the performance of using a series of hand-crafted noise ratios to guide the label denoising procedure. These manually set noise ratios contain no prior information. As shown in Table 2a, our noise estimator provides sound guidance in determining noisy labels.

Study Thresholds for Noise Estimation. We study the impact of thresholds, *i.e.*, θ^a and θ^v, in noise ratio estimation. A lower threshold leads to smaller noise ratios and vice versa. Since the predictions of the noise estimator are normalized by the mean value per category, θ^a (θ^v) may be more than 1. We find in Table 2b that our method is robust when θ^a and θ^v are within a reasonable range.

Study the Impact of Cross-modal Attention on Noise Estimator. As shown in Table 2c, it leads to a noticeable performance drop when training the noise esti-

Table 2: **Ablation studies.** Tables 2a, 2b, and 2c are studies on estimating noise ratios. Tables 2d, 2e and 2f are ablations on modality-specific label denoising.

(a) **Study the effectiveness of noise ratio estimator**. The constant noise ratios are hand-crafted.

Noise Ratio	A	V	AV	Type	Event
0.1	60.9	53.9	51.4	55.4	55.4
0.2	61.3	54.4	52.0	55.9	55.9
0.3	60.8	55.2	51.6	55.9	56.1
0.4	60.2	56.6	53.0	56.6	56.2
0.5	58.4	58.3	53.2	56.6	55.9
Estimated ratios	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**

(b) **Study thresholds for noise estimation.** Segment-level Type@AV results are reported.

audio θ^a \ visual θ^v	1.6	1.7	1.8	1.9	2.0
0.5	58.3	58.9	60.6	60.5	60.0
0.6	58.9	58.8	**60.8**	60.5	60.5
0.7	58.5	59.0	60.8	60.7	60.4
0.8	58.3	59.3	60.7	60.6	60.4

(c) **Study the impact of cross-modal attention on noise estimator**. "cm attn." represents cross-modal attention.

Estimator	A	V	AV	Type	Event
$w/$ cm attn.	60.9	55.9	52.9	56.6	56.3
w/o **cm attn.**	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**

(d) **Intra-modal label denoising *vs.* Joint-modal label denoising**. "In-MoLD" indicates Intra-modal label denoising.

Methods	A	V	AV	Type	Event
InMoLD	59.6	58.5	51.8	56.6	57.8
JoMoLD	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**

(e) **Single-modal label denoising *vs.* joint-modal label denoising.** "Audio only" or "Visual only" denotes that label denoising is performed only for audio or visual track.

Modality	A	V	AV	Type	Event
Audio only	61.3	53.2	50.3	54.9	56.2
Visual only	61.0	62.7	56.2	60.0	59.2
both (JoMoLD)	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**

(f) **Study the warm-up of noise ratios.** The noise ratios will be increased from 0 to its real values during the period of warm-up.

Warm-up epochs	A	V	AV	Type	Event
no warm-up	60.6	63.3	56.3	60.1	59.0
0.7	61.1	63.6	56.8	60.5	59.7
0.8	60.9	63.8	56.6	60.4	59.5
0.9	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**
1.0	61.3	63.7	56.8	60.6	59.7

mator with cross-modal attention. Since cross-attention mixes the clues between modalities, it causes improper noise estimation for each modality. Experiments also verify the rationality of removing cross-attention in the noise estimator.

Ablation Studies on Modality-Specific Label Denoising:

Intra-modal Label Denoising vs. Joint-modal Label Denoising. Experiments in Table 2d demonstrates the superiority of JoMoLD over intra-modal label denoising (InMoLD). The latter does not consider cross-modal clues for label denoising. Specifically, for audio modality, InMoLD only considers the labels of samples with high losses in audio modality as audio noise, and does not check whether the losses in the visual modality of these samples are low. The procedure is the same for visual modality label denoising. JoMoLD makes good use of the intuition that a label would not serve as noise for both modalities, so a confident noisy label should enjoy the different loss patterns between two modalities.
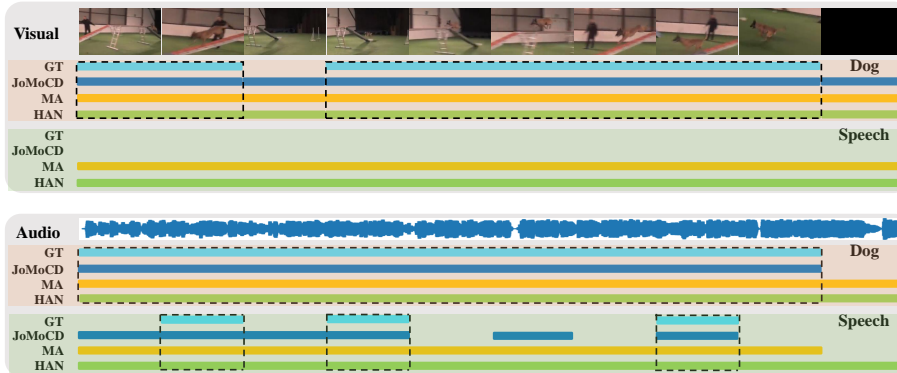
Fig. 3: **Qualitative comparisons with the state-of-the-art mothods**. We detail the parsing visualization results of "Dog" and "Speech" categories for visual and audio modalities. "GT" denotes the ground truth annotations. We compare JoMoLD, MA [43] and HAN [37] with GT. Generally, JoMoLD achieves better parsing results.

Single-modal Label Denoising vs. Joint-modal Label Denoising. Single-modal label denoising adopts the same method as joint-modal label denoising, but conducts label denoising only for a single modality. Table 2e shows that denoising labels for the visual track brings more improvement than the audio track. This might be because LLP is an audio-dominant dataset with more visual noise. Moreover, removing noisy labels for both modalities brings further improvement.

Study Warm-up Strategy. At the beginning of the training, the model treats clean and noisy labels alike, so we cannot distinguish them from the losses. In order to avoid selection bias in the early training, we adopt a warm-up strategy. It gradually increases the noise ratios from zero to the pre-computed values during warm-up epochs. The results are shown in Table 2f. Because warm-up strategy is not a critical step in JoMoLD, we omit it in Algorithm 1 for clarity.

4.4   Qualitative Analysis

Video Parsing Visualization. Figure 3 visualizes the parsing results of JoMoLD, MA [43] and HAN [37] as well as the ground truth annotation "GT". This video contains audio events "Dog" and "Speech", and visual event "Dog". Our method achieves the best performance in event recognition and localization among three methods. In detail, MA and HAN wrongly predict the "Speech" event in the visual track while JoMoLD correctly predicts no "Speech" event in the visual track. This can be credited to our more accurate label denoising procedure during training. For the audio track, JoMoLD makes more precise detection results than the other two for the event "Dog" and "Speech". Nevertheless, our method still makes mistakes for some segments. Due to the lack of segment-level supervised signals, there is still room for our method to improve performance.
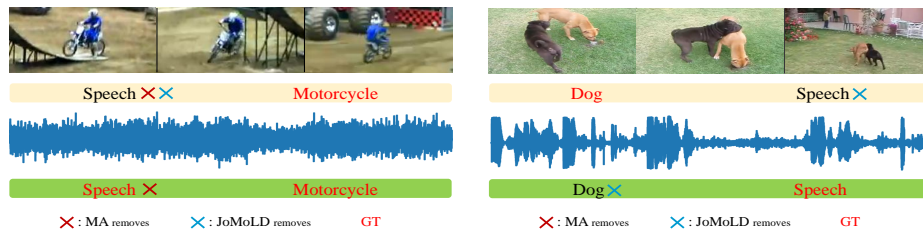
(a) JoMoLD correctly removes the label "Speech" for visual track and remains it for audio track. But MA fails to recognize audio clues of "Speech" and removes it.

(b) There are no "Dog" clues appear in audio track and no "Speech" clues appear in visual track. MA fails to identify the noise while JoMoLD correctly removes them.

Fig. 4: Comparisons of label denoising results between JoMoLD and MA.

Label Denoising Visualization. We visualize two cases of label denoising results of JoMoLD and MA [43] in Figure 4. The first case displays a motorcycle race. The event "Motorcycle" can be perceived from both modalities, but "Speech" is from the off-screen audience. In the second case, the event "Dog" only appears in the visual track and "Speech" in the audio track. MA makes mistakes in both two cases. When identifying noisy labels, MA exchanges the audio tracks of one video with another video whose label sets do not intersect. So the modality predictions of the newly assembled video are lowered when cross-attended to the unrelated modality. While our method does not confuse the video content, and successfully removes noisy labels while retaining correct labels.

More visualizations are presented in the appendix.

5   Conclusions

In our work, we focus on weakly-supervised audio-visual video parsing task. We are committed to solving the modality-specific label noise issue, which degenerates parsing performance according to our analysis. We notice that the clean and noisy labels present different loss patterns, and an annotated event label would not be noise for both modalities. Thus we take the different loss levels of the two modalities as the noise and remove the noisy labels. Extensive experiments show that our Joint-Modal Label Denoising method selects modality-specific noise more accurately and improves performance over the state of the arts. As for the limitations, more large-scale datasets of the weakly-supervised AVVP task are expected to further validate our method in future work.

References

1. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020). NeurIPS (2020) 3, 4
2. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6644–6652 (2021) 5
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 1, 3, 4
4. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018) 4
5. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International Conference on Machine Learning. pp. 312–321. PMLR (2019) 4
6. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017) 3, 4, 9
7. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. Advances in neural information processing systems **29**, 892–900 (2016) 3, 4
8. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998) 4
9. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4), 1–11 (2018) 3
10. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15495–15505 (2021) 3
11. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020) 4
12. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2827–2836 (2016) 3, 4
13. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I.W., Sugiyama, M.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 8536–8546 (2018) 4
14. 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
15. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017) 10

16. Hori, C., Hori, T., Lee, T.Y., Zhang, Z., Harsham, B., Hershey, J.R., Marks, T.K., Sumi, K.: Attention-based multimodal fusion for video description. In: Proceedings of the IEEE international conference on computer vision. pp. 4193–4202 (2017) 4

17. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9248–9257 (2019) 1

18. Hu, P., Peng, X., Zhu, H., Zhen, L., Lin, J.: Learning cross-modal retrieval with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5403–5413 (2021) 5

19. Kim, Y., Yun, J., Shon, H., Kim, J.: Joint negative and positive learning for noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9442–9451 (2021) 4

20. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 7774–7785 (2018) 3, 4

21. Li, J., Socher, R., Hoi, S.C.H.: Dividemix: Learning with noisy labels as semi-supervised learning. ArXiv **abs/2002.07394** (2020) 2, 4

22. Li, M., Soltanolkotabi, M., Oymak, S.: Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: The 23rd International Conference on Artificial Intelligence and Statistics (2020) 3, 4, 9

23. 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) 4, 10, 11

24. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019) 10, 11

25. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in Neural Information Processing Systems **33** (2020) 4

26. Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 961–971 (2017) 4

27. Mandal, D., Bharadwaj, S., Biswas, S.: A novel self-supervised re-labeling approach for training with noisy labels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1381–1390 (2020) 4

28. 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) 10, 11

29. Nguyen, T., Mummadi, C., Ngo, T., Beggel, L., Brox, T.: Self: learning to filter noisy labels with self-ensembling. In: International Conference on Learning Representations (ICLR) (2020) 2, 4

30. Panda, R., Chen, C.F., Fan, Q., Sun, X., Saenko, K., Oliva, A., Feris, R.: Adamml: Adaptive multi-modal learning for efficient video recognition. arXiv preprint arXiv:2105.05165 (2021) 1, 4

31. Pedro Morgado, Nuno Vasconcelos, I.M.: Audio-visual instance discrimination with cross-modal agreement. In: Computer Vision and Pattern Recognition (CVPR), IEEE/CVF Conf. on (2021) 1

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

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

34. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. pp. 5907–5915. PMLR (2019) 4

35. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5552–5560 (2018) 4

36. 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

37. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 436–454. Springer (2020) 2, 4, 5, 6, 8, 9, 10, 11, 13

38. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018) 1, 10, 11

39. 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

40. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 322–330 (2019) 4

41. 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) 10, 11

42. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13726–13735 (2020) 2, 10, 11

43. Wu, Y., Yang, Y.: Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1326–1335 (2021) 2, 4, 6, 7, 10, 11, 13, 14

44. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6292–6300 (2019) 4

45. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020) 1, 4

46. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7017–7025 (2019) 4

47. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning. pp. 7164–7173. PMLR (2019) 2, 4, 10, 11

48. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization (2016). arXiv preprint arXiv:1611.03530 (2017) 3, 4, 9

49. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: 32nd Conference on Neural Information Processing Systems (NeurIPS) (2018) 4

50. 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) 3

51. Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z.: Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: European Conference on Computer Vision. pp. 52–69. Springer (2020) 3

52. Zhou, J., Zheng, L., Zhong, Y., Hao, S., Wang, M.: Positive sample propagation along the audio-visual event line. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8436–8444 (2021) 4