Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing

Yu Wu\textsuperscript{1,2}, Yi Yang\textsuperscript{2}\textsuperscript{*}
\textsuperscript{1}Baidu Research \textsuperscript{2}ReLER, University of Technology Sydney
yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au

Abstract

We investigate the weakly-supervised audio-visual video parsing task, which aims to parse a video into temporal event segments and predict the audible or visible event categories. The task is challenging since there only exist video-level event labels for training, without indicating the temporal boundaries and modalities. Previous works take the overall event labels to supervise both audio and visual model predictions. However, we argue that such overall labels harm the model training due to the audio-visual asynchrony. For example, commentators speak in a basketball video, but we cannot visually find the speakers. In this paper, we tackle this issue by leveraging the cross-modal correspondence of audio and visual signals. We generate reliable event labels individually for each modality by swapping audio and visual tracks with other unrelated videos. If the original visual/audio data contain event clues, the event prediction from the newly assembled data would still be highly confident. In this way, we could protect our models from being misled by ambiguous event labels. In addition, we propose the cross-modal audio-visual contrastive learning to induce temporal difference on attention models within videos, i.e., urging the model to pick the current temporal segment from all context candidates. Experiments show we outperform state-of-the-art methods by a large margin.

1. Introduction

We humans explore and perceive the sounding environments with sensory streams, including visual, auditory, tactile, etc. Among these simultaneous sensory streams, vision and audio are two fundamental streams that widely convey massive information in our daily life.

Audio-visual comprehension \cite{23, 40, 9, 39, 50} is more robust in identifying the ongoing events compared to those vision models \cite{45, 34}. For example, occlusions and blind spots are common in egocentric videos and web videos, where the object of interest is outside of the field-of-view (FoV). In such situations, auditory signals could provide reliable clues for video understanding.

Existing audio-visual research works \cite{1, 5, 7, 10, 12, 19, 6, 21, 27, 31, 53, 54, 57} usually assume audio and visual data are always correlated and temporally aligned. However, this alignment might not always hold in practice. We may find lots of videos whose sound originates outside of the scene view. Despite the nonalignment, audio signals are still important in understanding the events, such as out-of-screen motorcycle racing. In this paper, we focus on the audio-visual video parsing (AVVP) task \cite{39}, which aims at providing a detailed analysis of auditory, visual, and audio-visual events in videos without such alignment assumptions. As shown in Fig. 1, the target of AVVP is to recognize event categories in each sensory modality and localize them temporally in videos.

Due to exhausting labeling cost, Tian \textit{et al.} \cite{39} proposed the weakly-supervised learning for the AVVP task, which only requires sparse labeling on the presence or absence of
event categories for training. The weakly supervised labels only indicate which event occurs in the video, without detailed modalities and temporal boundaries. The weak labels are more comfortable to annotate and can be boosted with automatic annotation (tags) for web videos. To solve the challenging issue, Tian et al. [39] proposed introducing cross-modal and self-modal attention to obtain aggregated features. The model is optimized in the Multimodal Multiple Instance Learning (MMIL) way, which regards overall event labels as the optimization targets for both audio and visual predictions.

However, audio and visual content are naturally different sensory streams. Visual data are captured by specific camera views, while audio signals collected by microphones could perceive all audible events of the scenes. Unlike other weakly supervised learning tasks, some event information may only exist in a single modality (either audio signals or visual signals). It would be irrational to optimize both modality predictions to be close to the overall event labels.

In this paper, we propose to tackle the challenging task by exploring heterogeneous clues. We alleviate the modality uncertainty issue and generate reliable event labels individually for each modality without additional annotations. To achieve the goal, we exchange the audio and visual track of a training video with other unrelated videos. Our motivation is that the newly assembled video’s prediction would still be highly confident if the visual/audio signals do contain clues of the target event. Otherwise, the event information is not visible/audible in the corresponding modality. In this way, we could obtain precise modality-aware event labels and protect models from being misled by the ambiguous overall labels. To the best knowledge of ours, we are the first that swap audio and visual tracks with other videos to assess the modality uncertainty.

In addition, we also propose to induce temporal difference within videos in a contrastive learning manner. Previous methods obtain enhanced modality features by leveraging all temporal contexts of the whole video. We argue that these might harm the model performance since it obscures the temporal difference within an event video. Since we do not have temporal annotations in training, inspired by self-supervised learning [16, 51], we propose to introduce contrastive learning to introduce temporal difference into aggregated features. We urge the attention model to pick the correct temporal cross-modal segment features from all candidate distractors. Thus the aggregated feature would be more likely the information that happens at this segment instead of all context features, leading to better temporal localization performances. To summarize, our contributions are as follows:

- We propose to address the modality uncertainty issue by exchanging audio and visual tracks with other videos. Thus we can obtain accurate modality-aware event supervision instead of ambiguous overall labels.
- We further introduce temporal heterogeneous constraint into the attention model via contrastive learning, which alleviates the ambiguous temporal boundaries issues in the weakly-supervised AVVP task.
- Experiments show our method significantly outperforms the state-of-the-art methods by a large margin on all evaluation metrics. Specifically, we improve the segment-level audio-visual parsing accuracy from 48.9% to 55.1% on the LLP dataset.

2. Related Work

We first discuss the joint modeling of audio-visual modalities, and then discuss the temporal action localization for video understanding. Finally, we discuss our focus’s related progress, the audio-visual video parsing, and the event localization problem.

2.1. Audio-Visual Representation Learning

Many works focus on joint learning for vision and audio signals. Most works [1, 2, 27, 28, 21, 58] assume that audio and visual data are synchronized and thus treat the audio and visual learning in a self-supervision way. Aytar et al. [2] propose SoundNet by designing a visual teacher network to learn audio representations from unlabeled videos. Owens et al. [28] leverage ambient sounds as supervision to learn visual representations. Arandjelovic and Zisserman [1] propose to learn both visual and audio representations in an unsupervised manner through an audio-visual correspondence task. Some works [21, 27] learn such visual and audio representation by the audio-visual temporal synchronization task. Besides audio-visual representations learning, there are many audio-visual applications such as sound separation [5, 7, 10, 11, 12, 53, 54, 57], sound source localization [27, 31, 40], audio-visual action recognition [13, 20], audio-visual navigation [3, 8], audio-visual video captioning [30, 37, 38, 46], and audio-visual event localization [23, 40, 41, 50].

2.2. Video Understanding and Action Localization

Deep learning methods have achieved promising performance in understanding video content [35, 45, 47]. Simonyan et al. [35] proposed Two-Stream to utilize both RGB frames and optical flow as the 2D CNN input to modeling appearance and motion, respectively. Temporal Segment Networks (TSN) [45] extended the two-stream CNN by extracting features from multiple temporal segments.
Tran et al. [42] proposed a 3D CNN to learn the spatio-temporal information.

Different from the action recognition task, action localization [34, 22, 32, 56, 25, 55, 52] aims at localizing actions within untrimmed videos. Previous supervised methods for action localization [32, 34, 56] usually first generate action proposal candidates and then predict the action based on these proposals. The proposal-classification methods usually filter out the background frames at the proposal stage via a binary actionness classifier.

There are also weakly-supervised works [24, 26, 29, 36, 44] proposed for action localization. These methods usually use Multiple Instance Learning (MIL) for training without temporal boundary annotations. Wang et al. [44] proposed UntrimmedNet composed of a classification module and a selection module. Nguyen et al. [26] introduced a sparsity regularization for video-level classification. Shou [33] explored score contrast in the temporal dimension for weakly supervised localization. Unlike the weakly supervised action localization task, we focus on localizing events in audio-visual video parsing, which contains motionless or even out-of-screen sound sources.

2.3. Audio-Visual Video Parsing

Audio-visual video parsing [39] (AVVP) aims at providing a detailed analysis of auditory and visual events in videos. It parses unconstrained videos into a set of video events associated with event categories, boundaries, and modalities. Early related works [23, 37, 50] focus on a similar task, i.e., audio-visual event localization, which localizes a visible and audible event/action in a video. AVE [40] is an audio-guided visual attention mechanism to adaptively learn which visual regions to look for the corresponding sounding object or activity. Lin et al. [23] propose integrating audio and visual features to a global feature in a sequence-to-sequence manner. Wu et al. [50] leverage the global event feature as the reference when localizing an event. In [39], Tian et al. propose a hybrid attention network and optimize the model in the Multi-modal Multiple Instance Learning (MMIL) way, i.e., taking the overall video-level label as the optimization targets for both audio and visual model predictions. However, we argue that such overall labels harm the model training due to the audio-visual asynchrony. Different from these methods, we generate reliable event labels individually for each modality to protect models from being misled by the ambiguous overall labels. In addition, we also induce temporal differences among segments by audio-visual contrastive learning.

3. Method

In this section, we introduce our method in detail. We begin with the preliminaries of the problem statement and introduce the baseline framework for this task. Then we illustrate the modality-aware event label refinement and our contrastive learning for audio-visual video parsing.

3.1. Preliminaries

Problem statement. In the AVVP task, each video may contain multiple visible or audible events. For a $T$-seconds audio-visual video sequence $S = \{V_i, A_i\}_{T=1}^T$, $A_i$ is the audio track and $V$ is the visual counterpart at the $t$-th segment. Each segment lasts for one second long. For evaluation, the targets are to predict the event labels for each segment and each modality. For the $t$-th video segment ($V_t, A_t$), the target $y_t = (y_{t}^{a}, y_{t}^{v}, y_{t}^{av})$ is a multi-class event label. Note there may exist zero or many events that are happening at the $t$-th moment. $y_{t}^{a}$, $y_{t}^{v}$, and $y_{t}^{av}$ are audio, visual, and audio-visual event labels, respectively. The audio-visual events $y_{t}^{av}$ occur only when events are both audible and visible at the same time.

For training, we only have access to weakly-supervised labels. Specifically, we only know events that show up in the video sequence $S$, but do not have precise labels such as the events occurring time and modalities. Therefore, the temporal and multi-modal uncertainty in the weakly-supervised AVVP problem makes it very challenging.

Data process. Pre-trained audio and visual deep models are applied to obtain visual representations $\{f_t^v\}_{T=1}^T$ and audio representations $\{f_t^a\}_{T=1}^T$ at the segment level (one second per segment), respectively. Following [39, 40, 50], the local feature extractor is fixed, and we build our method on top of these local features. The extracted audio and visual features are used as input for the following modeling.

Feature aggregation. Previous work [39] proves the effectiveness of feature aggregation upon the local input features. Thus we also enhance the input features by leveraging context information via self-attention and cross-attention mechanism. Denote $\text{att}(\cdot)$ to be the scaled dot-product conducted on the query, keys, and values,

$$\text{att}(q, K, V) = \text{Softmax}(\frac{qK^T}{\sqrt{d}})V,$$

where $d$ is the dimensionality of the feature vector $q$. The aggregated feature can be obtained by,

$$\hat{f}_t^a = f_t^a + \text{att}(f_t^a, F^a, \alpha^a) + \text{att}(f_t^a, F^v, \alpha^v),$$

$$\hat{f}_t^v = f_t^v + \text{att}(f_t^v, F^v, \alpha^v) + \text{att}(f_t^v, F^a, \alpha^a),$$

where $F^a = (f_1^a, ..., f_T^a)$ and $F^v = (f_1^v, ..., f_T^v)$ are the audio and visual features sequence from the same video $S$, respectively. Compared to the original input features, the aggregated features $\hat{f}_t^a$ and $\hat{f}_t^v$ are promoted by gathering event information across the entire video content.

Multiple Instance Learning. The event prediction of each segment and modality is based on the aggregated features.
Since there might be multiple events happening at the same segment, we use a Sigmoid function on the classifier to output probability for each event category. We denote $p^a_i$ and $p^v_i$ to be the event predictions on the audio and visual features at the $i$-th segment, respectively. However, we could only access a video-level weak label $\bar{y}$ instead of accurate audio and visual segment-level labels in the weakly-supervised training. Following [39], we use the attentive MIL pooling method to predict video-level event probability. Specifically, the video-level event probability $\bar{p}^a$ and $\bar{p}^v$ are obtained by the weighted average of all segment-level predictions. For our baseline, we optimize the video-level probability $\bar{p}^a$ and $\bar{p}^v$ to be close to the overall event labels $\bar{y}$ using the binary cross-entropy loss function.

### 3.2. Exchanging Audio and Visual Tracks

The above baseline could be used to train a decent model for weakly-supervised AVVP. However, it may induce severe label noise due to the modality uncertainty. Many events may only exist in one modality (either audio signals or visual signals) since audio and visual content are naturally different information sources. Optimizing both modality predictions ($i.e., \bar{p}^a$ and $\bar{p}^v$) to be close to the overall labels would inevitably introduce noise in training.

Motivated by the natural correlation between audio and visual content, we propose alleviating the modality uncertainty issue by exchanging audio and visual tracks with other videos. As shown in Fig. 2, we first assess modality uncertainty and then generate modality-aware event labels for each modality individually. Finally, we re-train our model from scratch based on these refined labels.

**Exchanging channels.** Our target is to localize the target event between modalities, $i.e.,$ whether a modality contains the target events or not. To achieve the goal, we leverage other videos to assess the target video without requiring additional annotations. Suppose we have two audio-visual videos that have disjoint video-level event labels, $i.e., S^i = (V^i, A^i)$ and $S^j = (V^j, A^j)$, but $\bar{y}^i \neq \bar{y}^j$. Taking the video $S^i = (V^i, A^i)$ as our target video, we exchange the visual channel and audio tracks of these two videos and form a new “video” by,

\[
\hat{S}^j = (V^j, A^i), \tag{4}
\]
\[
\hat{S}^i = (V^j, A^j), \tag{5}
\]

where $\hat{S}^j$ denotes the new “video” formed by the visual content from the video $S^i$ and the audio track from the video $S^j$. Since the video-level event labels $\bar{y}^j$ guarantee there is no event $\bar{y}^j$ existing in any modality of video $S^j$, we could safely conclude that both $V^j$ and $A^j$ are unrelated to the
target event $y^i$. Thus for the newly assembled data $\hat{S}^i_1$ and $\hat{S}^i_2$, the only clues about the event information $y^i$ are from the content of $i$-th video $S^i$, i.e., either from $V^i$, $A^i$ or both.

Assessing modality uncertainty. We assume that the newly assembled video’s prediction would still be highly confident if the visual/audio signals do contain clues of the target event. In other words, the event information is likely to be missed in the remaining modality if the prediction is low on the assembled videos. Denote the base model to be $\phi(\cdot)$, we obtain the event predictions for these assembled videos by,

$$p_{\hat{a}}^v, p_{\hat{a}}^a = \phi(V^i, A^i) / E_c,$$  \tag{6}

$$p_{\hat{v}}^v, p_{\hat{v}}^a = \phi(V^i, A^i) / E_c,$$  \tag{7}

where $p_{\hat{a}}^v$ indicates the event prediction based on aggregated visual features for the video with changed audio, and $p_{\hat{a}}^a$ means the event prediction based on aggregated visual features for the video with changed vision. $E_c$ is the normalized error rate of the target event category $c$ according to training predictions. The intuition is that the misaligned labels are more likely to happen if we found it hard to optimize the corresponding event categories (training accuracy on event category $c$ is lower). We believe the predictions $p_{\hat{a}}^v$ and $p_{\hat{a}}^a$ indicate the reliability of event labels for the visual track in video $S^i$. Similarly, $p_{\hat{v}}^v$ and $p_{\hat{v}}^a$ are used to validate the reliability of event labels for the audio track.

Refining modality-aware event labels. By assessing each modality’s confidence, we could further refine the event labels and have different event labels for the two modalities. We reassign the event label and remove unrelated labels for each modality if the confidences are lower than a threshold 0.5, since the sigmoid prediction ranges from 0 to 1. Specifically, we would discard the labels for visual modality if $p_{\hat{a}}^v < 0.5$ and $p_{\hat{a}}^a < 0.5$. Similarly, we would also remove the event labels for audio modality if $p_{\hat{a}}^v < 0.5$ and $p_{\hat{a}}^a < 0.5$. We could roughly estimate whether the event happens visually or audibly through modality-aware labels.

3.3. Learning Temporal Heterogeneous Clues

We further induce the temporal difference in the attention model. Although the self-modality and cross-modality attention (Eqs. (2) and (3)) lead to a more comprehensive understanding by leveraging audio-visual contexts, however, we argue that these might harm the model performance since it obscures the temporal difference within an event video. It is necessary to introduce the temporal difference during the weakly-supervised training.

Since we do not have temporal annotation for each segment, we propose to leverage contrastive learning to alleviate the issue. Contrastive learning [4, 49] is popular in self-supervised learning. We design a proxy task that urges the attention model to pick the correct temporal segment from all distractor segments, which prevents the aggregated model from being dominated by a few segment features.

We use Noise Contrastive Estimation (NCE) [15, 16, 51] to encourage the aggregated feature $f^i_a$ to be close to the low-level visual feature $f^i_v$ at the same timestamp, while is far away from visual features at other temporal segments. Thus, the only positive target is the ground truth feature $f^i_v$. We then build a set of candidates as distractors containing the same video’s visual features but at different time steps, i.e., $f_{t'}^i$ where $t' \neq t$. These candidates are hard to distinguish since they are very close to the ground truth frame feature $f^i_v$.

With the positive target and these distractors, we can add auxiliary supervision to the model with contrastive learning. We first calculate the cosine similarity between the predicted feature and the candidates, $f_{t'}^i f^i_a$. Here we enforce all vectors to be L2-normalized feature embeddings, i.e., $\|f^i_v\| = 1, \|f^i_a\| = 1$. Thus we have the following objective function at the time step $t$,

$$L_c = -\log \frac{\exp(f^i_v f^i_a / \tau)}{\sum_{j} \exp(f_{t'}^i f^i_a / \tau)}$$  \tag{8}

where $\tau$ is a temperature parameter that controls the concentration level of the distribution. Higher $\tau$ leads to a softer probability distribution. We set $\tau = 0.2$ in our experiments.

By combining the binary cross-entropy loss and the above contrastive loss, the attention model may not be dominated by some temporal segments. The aggregated feature would be more likely the information that happens at this segment instead of all context features, leading to better temporal localization performances.

4. Experiments

4.1. Experiment Setup

The Look, Listen and Parse (LLP) Dataset [39] contains 11,849 YouTube video clips and 25 event categories. It covers a wide range of daily life scenes, including human activities, animal activities, music performances, and vehicle sounds. The detailed events categories, including man speaking, dog barking, playing guitar, and frying food etc., lasts 10 seconds with both audio and video tracks. There are 7,202 videos that contain events from more than one event categories and per video has averaged 1.64 different event categories. For the weakly-supervised AVVP task, there are 10,000 videos for training, containing weak labels only (video-level event annotations on the presence or absence of different video events). To evaluate AVVP performance, the 1,849 validation and test videos have fully annotated labels, i.e., individual audio and visual events with second-wise temporal boundaries.

Evaluation Metrics. We evaluate our method by parsing all types of events (audio, visual, and audio-visual...
events) under both segment-level and event-level metrics. F-scores are used as the metrics to evaluate the predictions. The segment-level metrics evaluate segment-wise event prediction performance. Besides segment-level performance, the event-level results are also reported to indicate the performance in real applications. For computing event-level F-score results, we extract events by concatenating consecutive positive snippets in the same event categories and compute the event-level F-score based on $mIoU = 0.5$ as the threshold. In addition, we also evaluate the overall audio-visual scene parsing performance of our method by computing aggregated results, i.e., “Type@AV” and “Event@AV”. Specifically, Type@AV computes averaged audio, visual, and audio-visual event evaluation results, while Event@AV computes the F-score considering all audio and visual events for each sample rather than directly averaging results from different event types.

**Implementation Details.** We use the same visual features and audio features as previous works for a fair comparison. We use both the ResNet-152 [17] model pre-trained on ImageNet and 18 layer deep R(2+1)D [43] model pre-trained on Kinetics-400 to extract visual representations. We decode videos at 8 fps and input each segment (lasting one second) to obtain the 2D and 3D visual features. We regard the concatenation of the two visual features as the low-level visual feature. For the audio signals, we use the VGGish network [18] pre-trained on AudioSet [14] to extract 128-D features. We use Adam optimizer to train the framework with a mini-batch size of 16 and a learning rate of $3 \times 10^{-4}$. We train 40 epochs and drop the learning by a factor of 10 after 10 epochs. Our training pipeline includes three stages. First, we optimize a base model for audio-visual scene parsing using MIL and our proposed contrastive learning. Second, we freeze the model and evaluate each video by swapping its audio and visual tracks with other unrelated videos. Finally, we re-train the model from scratch using modality-aware labels. We name the final model as “MA” to distinguish it from the base model.

### 4.2. Comparison with State-of-the-art Results

We compare our model MA with weakly-supervised sound detection method TALNet [48], temporal action localization methods STPN [26] and CMCS [24], and state-of-the-art audio-visual event parsing methods including AVE [40], AVSDN [23], and HAN [39]. All the models, including ours, are trained for fair comparisons using the LLP training dataset only, including the same training data and pre-processed audio/visual features.

Table 1 shows the performances of our method MA and state-of-the-art methods on the LLP test set. It can be seen from the table that our method outperforms the state-of-the-art methods by a large margin on all audio-visual video parsing subtasks for both the segment-level and event-level metrics. Specifically, on the audio-visual event prediction, our MA beats the state-of-the-art method HAN [39] by 6.2 points (from 48.9% to 55.1%) at the segment level, and 6.0 points (from 43.0% to 49.0%) at the event level. The most significant improvement is found for visual event parsing, which validates our motivation that previous methods are suffered from the ambiguous overall labels of invisible events. The comparison with the state-of-the-art methods demonstrates that our model is able to predict significantly better event categories with accurate temporal locations.

### 4.3. Ablation Studies

**Effectiveness of Modality-aware Refinement.** We conduct the ablation studies to show the effectiveness of the modality-aware refinement. As shown in Table 2, “Baseline + R” indicates the results of the model trained with modality-aware refinement. By leveraging clues between the audio and visual tracks and assigning different labels for the two modalities, we find the model performance gets significantly improved. Table 2 shows our model “Baseline + R” outperforms the baseline by about 4 points at audio-visual event parsing evaluation metrics. Specifically, for the visual event parsing, the model with the modality-aware refinement significantly improves the performance by 4.6 points (from 52.9% to 57.5%) at the segment-level predic-

<table>
<thead>
<tr>
<th>Event type</th>
<th>Methods</th>
<th>Segment-level</th>
<th>Event-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-visual</td>
<td>AVE [40]</td>
<td>35.4</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>AVSDN [23]</td>
<td>37.1</td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>HAN [39]</td>
<td>48.9</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td><strong>MA (Ours)</strong></td>
<td><strong>55.1 (+6.2)</strong></td>
<td><strong>49.0 (+6.0)</strong></td>
</tr>
<tr>
<td>Audio</td>
<td>TALNet [48]</td>
<td>50.0</td>
<td>41.7</td>
</tr>
<tr>
<td></td>
<td>AVE [40]</td>
<td>47.2</td>
<td>40.4</td>
</tr>
<tr>
<td></td>
<td>AVSDN [23]</td>
<td>47.8</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>HAN [39]</td>
<td>60.1</td>
<td>51.3</td>
</tr>
<tr>
<td></td>
<td><strong>MA (Ours)</strong></td>
<td><strong>60.3 (+0.2)</strong></td>
<td><strong>53.6 (+2.3)</strong></td>
</tr>
<tr>
<td>Visual</td>
<td>STPN [26]</td>
<td>46.5</td>
<td>41.5</td>
</tr>
<tr>
<td></td>
<td>CMCS [24]</td>
<td>48.1</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>AVE [40]</td>
<td>37.1</td>
<td>34.7</td>
</tr>
<tr>
<td></td>
<td>AVSDN [23]</td>
<td>52.0</td>
<td>46.3</td>
</tr>
<tr>
<td></td>
<td>HAN [39]</td>
<td>52.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td><strong>MA (Ours)</strong></td>
<td><strong>60.0 (+7.1)</strong></td>
<td><strong>56.4 (+7.5)</strong></td>
</tr>
<tr>
<td>Type@AV</td>
<td>AVE [40]</td>
<td>39.9</td>
<td>35.5</td>
</tr>
<tr>
<td></td>
<td>AVSDN [23]</td>
<td>45.7</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>HAN [39]</td>
<td>54.0</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td><strong>MA (Ours)</strong></td>
<td><strong>58.9 (+4.9)</strong></td>
<td><strong>53.0 (+5.3)</strong></td>
</tr>
<tr>
<td>Event@AV</td>
<td>AVE [40]</td>
<td>41.6</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td>AVSDN [23]</td>
<td>50.8</td>
<td>37.7</td>
</tr>
<tr>
<td></td>
<td>HAN [39]</td>
<td>55.4</td>
<td>48.0</td>
</tr>
<tr>
<td></td>
<td><strong>MA (Ours)</strong></td>
<td><strong>57.9 (+2.5)</strong></td>
<td><strong>50.6 (+2.6)</strong></td>
</tr>
</tbody>
</table>

Table 1. Comparisons with the state-of-the-art methods of the audio-visual video parsing task on the LLP test dataset. Note that we use the same input features as the compared methods.
Table 2. Ablation studies of the proposed modules. Audio-visual video parsing accuracy (%) are reported on the LLP test dataset. “C” denotes the proposed contrastive learning for temporal localization. “R” is our modality-aware refinement by exchanging audio and visual channels.

<table>
<thead>
<tr>
<th>Event type</th>
<th>Methods</th>
<th>Segment-level</th>
<th>Event-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-visual</td>
<td>Baseline</td>
<td>48.9</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>Baseline + C</td>
<td>49.7</td>
<td>43.8</td>
</tr>
<tr>
<td></td>
<td>Baseline + R</td>
<td>52.6</td>
<td>45.8</td>
</tr>
<tr>
<td></td>
<td>Baseline + C + R</td>
<td>55.1</td>
<td>49.0</td>
</tr>
<tr>
<td>Audio</td>
<td>Baseline</td>
<td>60.1</td>
<td>51.3</td>
</tr>
<tr>
<td></td>
<td>Baseline + C</td>
<td>61.9</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td>Baseline + R</td>
<td>59.8</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>Baseline + C + R</td>
<td>60.3</td>
<td>53.6</td>
</tr>
<tr>
<td>Visual</td>
<td>Baseline</td>
<td>52.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>Baseline + C</td>
<td>53.1</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>Baseline + R</td>
<td>57.5</td>
<td>54.4</td>
</tr>
<tr>
<td></td>
<td>Baseline + C + R</td>
<td>60.0</td>
<td>56.4</td>
</tr>
<tr>
<td>Type@AV</td>
<td>Baseline</td>
<td>54.0</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>Baseline + C</td>
<td>54.9</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>Baseline + R</td>
<td>56.6</td>
<td>50.8</td>
</tr>
<tr>
<td></td>
<td>Baseline + C + R</td>
<td>58.9</td>
<td>53.0</td>
</tr>
<tr>
<td>Event@AV</td>
<td>Baseline</td>
<td>55.4</td>
<td>48.0</td>
</tr>
<tr>
<td></td>
<td>Baseline + C</td>
<td>56.2</td>
<td>49.0</td>
</tr>
<tr>
<td></td>
<td>Baseline + R</td>
<td>56.6</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>Baseline + C + R</td>
<td>57.9</td>
<td>50.6</td>
</tr>
</tbody>
</table>

Table 3. Analysis of the modality-aware refinement. “Audio” and “Visual” indicate that we only refine labels for the audio modality and the visual modality, respectively. Segment-level audio-visual video parsing results are reported.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Audio</th>
<th>Visual</th>
<th>Audio-Visual</th>
<th>Type@AV</th>
<th>Event@AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio only</td>
<td>60.5</td>
<td>52.7</td>
<td>51.8</td>
<td>55.0</td>
<td>54.2</td>
</tr>
<tr>
<td>Visual only</td>
<td>60.4</td>
<td>59.0</td>
<td>53.5</td>
<td>57.9</td>
<td>57.1</td>
</tr>
<tr>
<td>Both</td>
<td>60.3</td>
<td>60.0</td>
<td>55.1</td>
<td>58.9</td>
<td>57.9</td>
</tr>
</tbody>
</table>

Table 4. Analysis on different \( \tau \) values used in contrastive learning (Eqn.(8)). Smaller \( \tau \) leads to sharper probability distribution. Segment-level audio-visual video parsing results are reported.

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>Audio</th>
<th>Visual</th>
<th>Audio-Visual</th>
<th>Type@AV</th>
<th>Event@AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>61.3</td>
<td>58.3</td>
<td>54.5</td>
<td>58.4</td>
<td>57.8</td>
</tr>
<tr>
<td>0.2</td>
<td>60.3</td>
<td>60.0</td>
<td>55.1</td>
<td>58.9</td>
<td>57.9</td>
</tr>
<tr>
<td>0.3</td>
<td>60.5</td>
<td>60.3</td>
<td>54.9</td>
<td>58.7</td>
<td>57.9</td>
</tr>
<tr>
<td>0.4</td>
<td>60.3</td>
<td>59.9</td>
<td>55.0</td>
<td>58.5</td>
<td>57.3</td>
</tr>
</tbody>
</table>

Analysis of Modality Bias in Refinement. We further uncover the effect of modality-aware refinement by looking into modalities. We conduct experiments including 1) only refining audio labels, 2) only refining visual labels, and 3) refining both modalities labels. The results are reported in Table 3. We can find the most significant improvement is brought by refining event labels for visual parsing prediction. By refining visual parsing labels, we significantly improve the performance on segment-level visual parsing evaluation. The reason is that the visual content could only be captured for specific camera views, whether the object of interest might usually be outside of the field-of-view. In contrast, the audio signals are collected by microphones, which are able to perceive all the event information of the scenes. Therefore, unmatched event labels are more common for visual modalities. By refining visual event labels for these **audible but not visible** videos, we observe a noticeable performance improvement on all the evaluation metrics except audio-only parsing.

Besides, we achieve further performance improvement by refining event labels for both modalities. Compared to “visual-only”, the model trained with both modality refine-

ment obtain considerable performance gain on all evaluation metrics.

Effectiveness of Cross-modal Contrastive Learning. Table 2 also shows the relative improvement brought by the cross-modal contrastive learning. Compared to the baseline, our model with the contrastive learning only (“Baseline + C”) shows an improvement on audio-visual even parsing. The relative improvement is even more significant when combining with the modality-aware refinement. By comparing the model “Baseline + C + R” and model “Baseline + R”, we can find the contrastive learning further improve the event parsing performance by about 2 points on most evaluation metrics. It indicates our proposed contrastive learning could introduce essential temporal differences for audio-visual video parsing.

Analysis of different \( \tau \) values. As indicated in Eqn.(8), \( \tau \) is a temperature parameter that controls the concentration level of the distribution. We validate different \( \tau \) values used in our experiments. Table 4 shows the comparison of the segment-level audio-visual video parsing evaluation. Smaller \( \tau \) leads to a sharper probability distribution. In experiments, we find the performances get slightly higher as \( \tau \) decreases. Overall speaking, our model is not sensitive to the values of \( \tau \) used in the contrastive learning (Eqn.(8)). In all other experiments, we set \( \tau \) to 0.2.

4.4. Qualitative Results

We visualize the audio-visual video parsing results in Fig. 3. “Pred” shows the prediction from our models. “GT” is the ground truth annotation. Overall speaking, our model could correctly recognize the events happening in the video. But it makes mistakes on the temporal location of these events. For example, our model still predicts guitar for the
Figure 3. Qualitative results on the LLP test set. The upper and bottom figure shows visual and audio event parsing, respectively. “Pred” is the prediction result from our model, while “GT” indicates the ground truth annotation.

<table>
<thead>
<tr>
<th>Pred:</th>
<th>GT:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guitar</td>
<td>Guitar</td>
</tr>
<tr>
<td>Singing</td>
<td>Singing</td>
</tr>
</tbody>
</table>

Figure 4. Examples of our refined labels for the visual modality.

<table>
<thead>
<tr>
<th>Fire alarm</th>
<th>Speech</th>
<th>Fire alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singing</td>
<td>Cheering</td>
<td>Singing</td>
</tr>
</tbody>
</table>

visual event parsing after 2s, although we could not find such clues of the guitar in the corresponding visual frames. The reason might be that the context feature aggregation collects too much information from the audio and video of other time stamps. For example, the audio clearly indicates “guitar” at this moment. Compared to the visual parsing, the audio event parsing prediction is more reliable in general. The reason might be that audio is more clear and easy to be distinguished compared to complex visual frames.

We also show two examples of our modality-aware label refinement in Fig. 4. By exchanging audio and visual tracks among training videos, we localized event clues and found some events do not exist in the visual/audio modality. The upper case in the figure is a news video about the fire alarm event. Although the event labels are “fire alarm” and “speech” for the entire video in training, the model does not predict the “speech” event given the assembled video with exchanged channels (consisting of the original visual content and a new audio track). Through exchanging audio and visual signals, we could obtain a more accurate event label for the visual modality, i.e., “fire alarm” only. In this way, we protect the visual model from being misled by the ambiguous overall event label “Speech”.

5. Conclusions

We focus on the weakly-supervised audio-visual video parsing task, which predicts the audible or visible event categories and their temporal locations. We believe it harms the model training if we train both audio and visual models using the same overall labels. We propose to generate modality-aware event labels by swapping audio and visual tracks with other unrelated videos. If the predictions on the new assembled data are not confident at the target event, there might be no events clues in the original visual/audio tracks. In this way, we could protect our models from being misled by ambiguous event labels. Besides, we further leverage heterogeneous clues temporally and induce temporal difference within videos by audio-visual contrastive learning. Experiments show we outperform state-of-the-art methods by a large margin. In conclusion, we found it useful by mining detailed annotations for different modalities. Inducing temporal difference also improves performance in the weakly-supervised AVVP task.

Acknowledgement. This research is in part supported by the ARC Discovery Project DP200100938.
References

[31] 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. 1, 2


