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

Yu Wu[1,2], Yi Yang[2*]
[1]Baidu Research [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 [23, 40, 9, 39, 50] is more robust in identifying the ongoing events compared to those vision models [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.

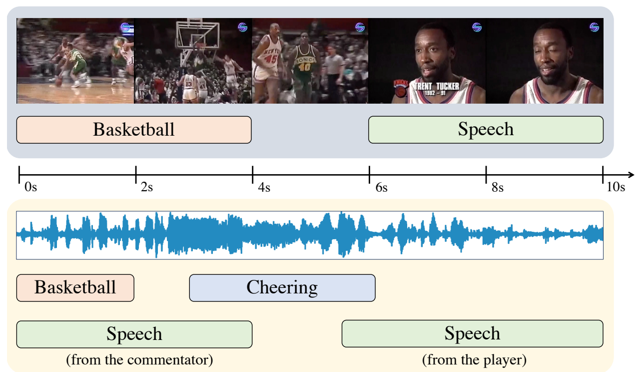

Figure 1. Examples of the audio-visual video parsing task. Colored rectangles indicate the ground truth events. Taking the visual and audio data as input, we aim at identifying the audible and visible events and their temporal location. Note that the visual and audio events might be asynchronous.

Existing audio-visual research works [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 [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 *et al*. [39] proposed the weakly-supervised learning for the AVVP task, which only requires sparse labeling on the presence or absence of

*This work was done when Yu Wu interned at Baidu Research. Yi Yang is the corresponding author.

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. For example, in a basketball match video, there might be commentators speaking, but we cannot find them visually (see Fig. 1). It harms the visual model optimization if we follow the universal weakly supervised learning way.

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 constrain 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 spatial-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 $\mathcal{S} = \{V_t, A_t\}_{t=1}^{T}$, A 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 $\mathbf{y}_t = (y_t^a, y_t^v, y_t^{av})$ 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 $\mathcal{S}$, 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 $\{\mathbf{f}_t^v\}_{t=1}^{T}$ and audio representations $\{\mathbf{f}_t^a\}_{t=1}^{T}$ 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 $\mathtt{att}(\cdot)$ to be the scaled dot-product conducted on the query, keys, and values,

$$\mathtt{att}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathtt{Softmax}(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{1}$$

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

$$\hat{\mathbf{f}}_t^a = \mathbf{f}_t^a + \mathtt{att}(\mathbf{f}_t^a, \mathbf{F}^a, \mathbf{F}^a) + \mathtt{att}(\mathbf{f}_t^a, \mathbf{F}^v, \mathbf{F}^v), \tag{2}$$

$$\hat{\mathbf{f}}_t^v = \mathbf{f}_t^v + \mathtt{att}(\mathbf{f}_t^v, \mathbf{F}^v, \mathbf{F}^v) + \mathtt{att}(\mathbf{f}_t^v, \mathbf{F}^a, \mathbf{F}^a), \tag{3}$$

where $\mathbf{F}^a = (\mathbf{f}_1^a, ..., \mathbf{f}_T^a)$ and $\mathbf{F}^v = (\mathbf{f}_1^v, ..., \mathbf{f}_T^v)$ are the audio and visual features sequence from the same video $\mathcal{S}$, respectively. Compared to the original input features, the aggregated features $\hat{\mathbf{f}}_t^a$ and $\hat{\mathbf{f}}_t^v$ 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.

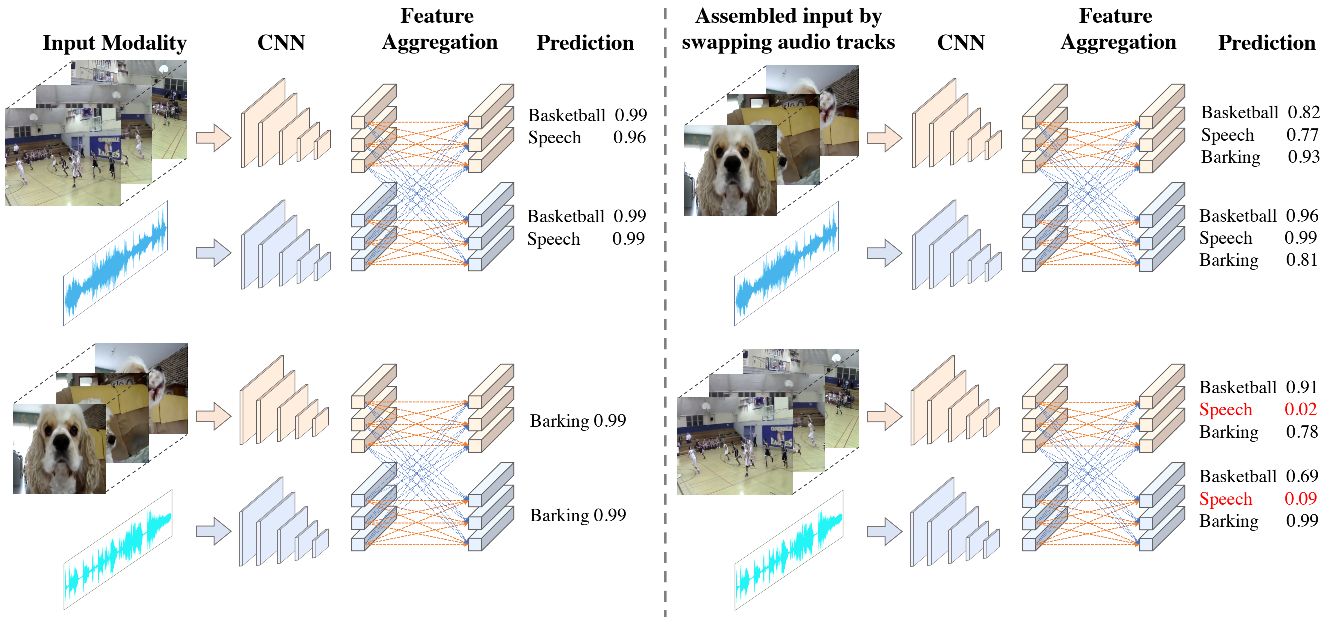

Figure 2. Our modality-aware label refining (MA) pipeline. Our model aggregates feature by self- and cross-modality attention, and then predicts the event labels for each modality. The figure's left shows the prediction on normal training videos, which would have relatively high confidence in their event predictions. We then exchange audio and visual tracks of these two unrelated videos (whose labels are not *disjoint*). The newly assembled videos are further input to the model for checking prediction confidences (right figure). We believe the confidences should still be high if the remaining visual/audio track does contain the target event. Otherwise, the event is not visible/audible in this modality. In this way, we could obtain modality-aware event labels and protect models from being misled by the ambiguous overall labels. In the case shown in the figure, we filter out the "Speech" event that is not visible in the original basketball video.

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_t^a and p_t^v to be the event predictions on the audio and visual features at the t-th segment, respectively. However, we could only access a video-level weak label $\bar{\mathbf{y}}$ 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.*, $\mathcal{S}^i = (V^i, A^i)$ and $\mathcal{S}^j = (V^j, A^j)$, but $\bar{\mathbf{y}}^i \neq \bar{\mathbf{y}}^j$. Taking the video $\mathcal{S}^i = (V^i, A^i)$ as our target video, we exchange the visual channel and audio tracks of these two videos and form a new "video" by,

$$\hat{\mathcal{S}}_j^i = (V^i, A^j), \tag{4}$$

$$\hat{\mathcal{S}}_i^j = (V^j, A^i), \tag{5}$$

where $\hat{S}_j^i$ denotes the new "video" formed by the visual content from the video $\mathcal{S}^i$ and the audio track from the video $\mathcal{S}^j$. Since the video-level event labels $\bar{\mathbf{y}}^j$ guarantee there is no event $\bar{\mathbf{y}}^i$ 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 $\bar{\mathbf{y}}^i$. Thus for the newly assembled data $\hat{\mathcal{S}}_j^i$ and $\hat{\mathcal{S}}_i^j$, the only clues about the event information $\mathbf{y}^i$ are from the content of i-th video $\mathcal{S}^i$, *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,

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

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

where $p_{\hat{a}}^v$ indicates the event prediction based on aggregated visual features for the video with *changed audio*, and $p_{\hat{v}}^v$ 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 event labels for *visual* modality if $p_{\hat{a}}^v < 0.5$ and $\hat{p}_{\hat{a}}^a < 0.5$. Similarly, we would also remove the event labels for *audio* modality if $\hat{p}_{\hat{v}}^v < 0.5$ and $\hat{p}_{\hat{v}}^a < 0.5$. 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 (Eqn. (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 $\hat{\mathbf{f}}_t^a$ to be close to the low-level visual feature $\mathbf{f}_t^v$ 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 $\mathbf{f}_t^v$. We then build a set of candidates as distractors containing the same video's visual features but at different time steps, *i.e.*, $\mathbf{f}_{t'}^v$ where $t' \neq t$. These candidates are hard to distinguish since they are very close to the ground truth frame feature $\mathbf{f}_t^v$.

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, $\mathbf{f}_j^{v\top}\hat{\mathbf{f}}_t^a$. Here we enforce all vectors to be L2-normalized feature embeddings, *i.e.*, $||\mathbf{f}_j^v|| = 1$, $||\hat{\mathbf{f}}_t^a|| = 1$. Thus we have the following objective function at the time step t,

$$\mathcal{L}_c = -\log \frac{\exp(\mathbf{f}_t^{v\top}\hat{\mathbf{f}}_t^a/\tau)}{\sum_j \exp(\mathbf{f}_j^{v\top}\hat{\mathbf{f}}_t^a/\tau)} \tag{8}$$

where τ is a temperature parameter that controls the concentration level of the distribution. Higher τ 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 3×10^{-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.

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

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.

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-

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

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.

tion and 5.5 points (from 48.9% to 54.4%) at the event level. It validates that ambiguous video-level labels harm model training since some events only appear in one modality.

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 refinement obtain considerable performance gain on all evaluation metrics.

Modality	Audio	Visual	Audio-Visual	Type@AV	Event@AV
Audio only	60.5	52.7	51.8	55.0	54.2
Visual only	60.4	59.0	53.5	57.9	57.1
Both	60.3	60.0	55.1	58.9	57.9

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.

τ	Audio	Visual	Audio-Visual	Type@AV	Event@AV
0.1	61.3	58.3	54.5	58.4	57.8
0.2	60.3	60.0	55.1	58.9	57.9
0.3	60.5	60.3	54.9	58.7	57.9
0.4	60.3	59.9	55.0	58.5	57.3

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

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

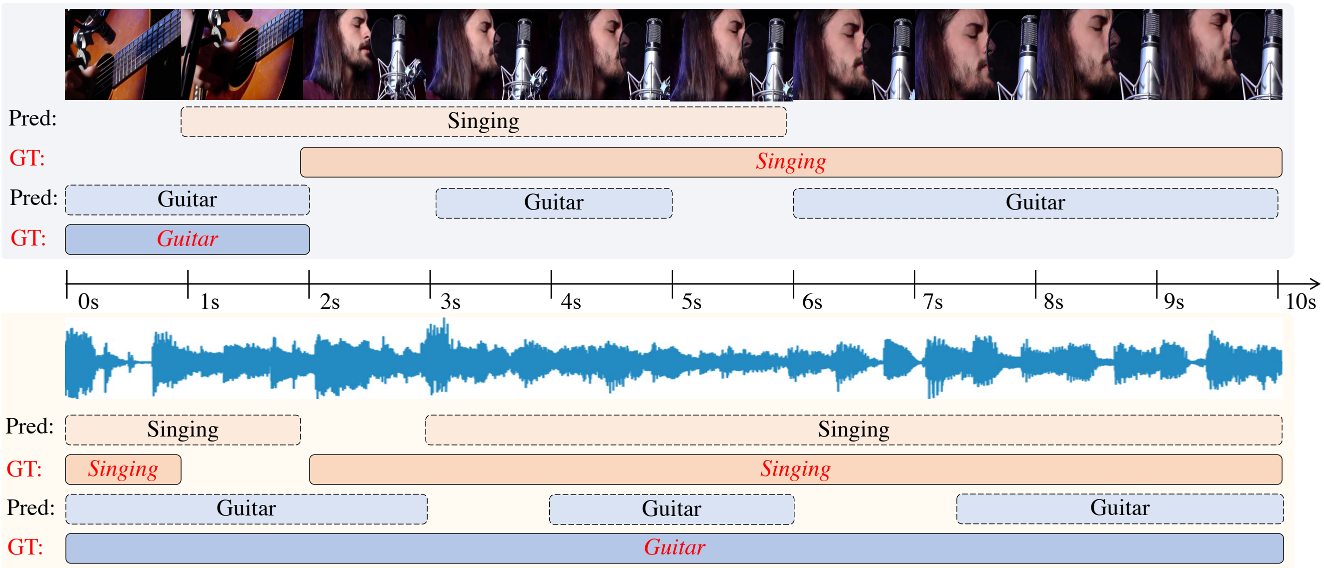

Figure 3. Qualitative results on the LLP test set. The upper and bottom figure shows visual and audio event parsing, respectively. "Pred" is the prediction result from our model, while "GT" indicates the ground truth annotation.

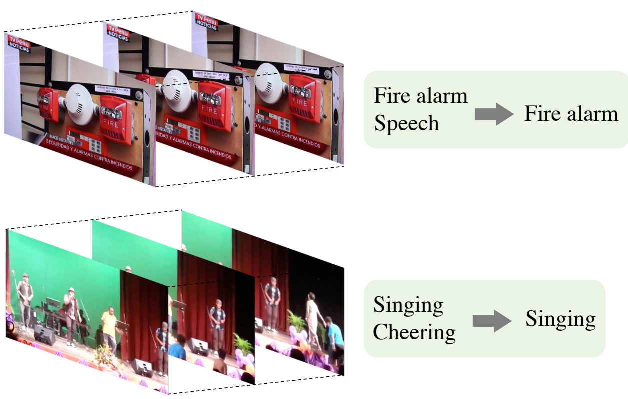

Figure 4. Examples of our refined labels for the visual modality.

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

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. *ICCV*, 2017. 1, 2

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016. 2

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICLR*, 2020. 5

[5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018. 1, 2

[6] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *ECCV*, 2020. 1

[7] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 1, 2

[8] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 2

[9] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, 2019. 1

[10] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 1, 2

[11] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, 2019. 2

[12] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1, 2

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

[14] 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

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 5

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 5

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

[18] 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

[19] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019. 1

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

[21] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 1, 2

[22] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, 2018. 3

[23] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*, 2019. 1, 2, 3, 6

[24] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019. 3, 6

[25] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, 2019. 3

[26] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 2018. 3, 6

[27] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1, 2

[28] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2

[29] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018. 3

[30] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, 2019. 2

[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

[32] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017. 3

[33] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, 2018. 3

[34] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 1, 3

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[36] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 3

[37] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872*, 2018. 2, 3

[38] Yapeng Tian, Chenxiao Guan, Goodman Justin, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *CVPR-W*, 2019. 2

[39] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6

[40] 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

[41] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in the wild. In *CVPR-W*, 2019. 2

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3

[43] 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

[44] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 3

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2

[46] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *NAACL*, 2018. 2

[47] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[48] Yun Wang, Juncheng Li, and Florian Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP*, 2019. 6

[49] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 2021. 5

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

[51] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2, 5

[52] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019. 3

[53] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 1, 2

[54] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2

[55] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, 2020. 3

[56] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 3

[57] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020. 1, 2

[58] Ye Zhu, Yu Wu, Hugo Latapie, Yi Yang, and Yan Yan. Learning audio-visual correlations from variational cross-modal generation. In *ICASSP*, 2021. 2