Rethink Cross-Modal Fusion in Weakly-Supervised Audio-Visual Video Parsing

Yating Xu Conghui Hu Gim Hee Lee
Department of Computer Science, National University of Singapore
xu.yating@u.nus.edu conghui@nus.edu.sg gimhee.lee@nus.edu.sg

Abstract

Existing works on weakly-supervised audio-visual video parsing adopt hybrid attention network (HAN) as the multi-modal embedding to capture the cross-modal context. It embeds the audio and visual modalities with a shared network, where the cross-attention is performed at the input. However, such an early fusion method highly entangles the two non-fully correlated modalities and leads to sub-optimal performance in detecting single-modality events. To deal with this problem, we propose the messenger-guided mid-fusion transformer to reduce the uncorrelated cross-modal context in the fusion. The messengers condense the full cross-modal context into a compact representation to only preserve useful cross-modal information. Furthermore, due to the fact that microphones capture audio events from all directions, while cameras only record visual events within a restricted field of view, there is a more frequent occurrence of unaligned cross-modal context from audio for visual event predictions. We thus propose cross-audio prediction consistency to suppress the impact of irrelevant audio information on visual event prediction. Experiments consistently illustrate the superior performance of our framework compared to existing state-of-the-art methods.

1. Introduction

With the ultimate goal of understanding both audio and visual content in video, multimodal video understanding finds a variety of applications in video retrieval [11], video surveillance [38] *etc*. As video is naturally equipped with both audio and visual signals, many prior works have incorporated audio modality into the analyses and shown its benefits to several emerging visual tasks [10, 14, 17, 30, 35]. Audio-Visual Video Parsing (AVVP) [34] is one of the most challenging tasks which aims at classifying and localizing the temporal event segments in the audio and visual streams respectively. The task requires the model to fully understand video content in both audio and visual streams while only video-level label is provided, as the fine-grained event labels for the two modalities are labour-intensive to source. Since there exists an intractable barrier to access full supervision, models must resort to a weakly-supervised paradigm by learning from the union of all events in a video without any modality and time indication.

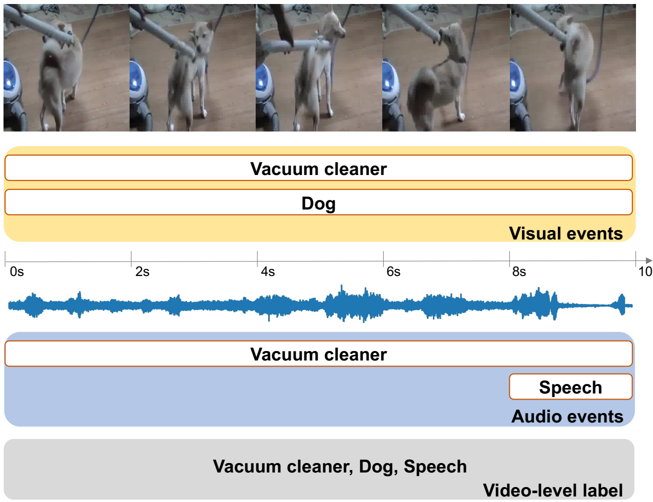

Figure 1. Illustration of audio-visual video parsing task. Given a video, it classifies events and detects their temporal locations in the audio and visual streams, respectively. During training, only video-level label is provided.

To address the challenging problem, the state-of-the-art approaches utilize the correlation between the audio and visual streams to guide the model training. For example, Hybrid Attention Network (HAN) [34] captures the cross-modal context by fusing the pre-extracted audio and visual features directly from the first layer of the network using cross-attention (as illustrated in Fig. 2(a)). However, this strong entanglement could undesirably mix uncorrelated information from audio and visual streams. Real scenes may include occlusion or may be captured by camera of limited field of view, which causes the audio and visual streams not fully correlated. For instance, in Fig. 1, the silent dog and off-camera human speech only appear in one single modality. Accordingly, we analyze the performance of HAN in detecting single-modality and multi-modality events

Model	Single-modality Event	Multi-modality Event
HAN	44.0	**67.2**
HAN-CA	47.0	58.8
Ours	**50.6**	66.1

Table 1. Analysis of HAN [34]. 'HAN-CA' denotes HAN without cross-attention. Segment-level evaluation is conducted.

as shown in Tab. 1. Single-modality events denote events *only* happening in audio or visual modality, while multi-modality events refer to events appearing with temporal overlap on audio and visual streams. The results are averaged F-scores per event. Compared with the original HAN, the HAN-CA[1], when excluding cross-modal fusion, exhibits a significant decrease in predicting multi-modality events. However, the prediction performance for single-modality events experiences an enhancement. It suggests that strong entanglement with another non-fully correlated modality is harmful in detecting its own exclusive events while the absence of fusion negatively affects the detection of audio-visual events. As either fully entangled fusion or complete independence of the two modalities can hurt the performance badly, it is imperative to design a better fusion strategy for the two partially correlated modalities.

To solve this problem, we propose messenger-guided mid-fusion transformer (MMT) to suppress the uncorrelated information exchange during fusion. Compared to the early fusion in HAN, the mid-fusion is more flexible in controlling the flow of the cross-modal context. It can first aggregate a clearer global understanding of the raw input sequence, which helps identify the useful cross-modal context in the fusion module. The messengers are the core of MMT, which serves as the compact cross-modal representation during fusion. (Fig. 2(b)). Due to their small capacity, they can help amplify the most relevant cross-modal information that agrees best with the clean labels while suppressing the noisy information that causes disagreement. Our MMT is able to largely improve the performance of detecting single-modality events while maintaining a relatively high performance of detecting multi-modality events.

We further propose cross-audio prediction consistency (CAPC) to suppress the undesired predictions in the visual stream caused by mismatched audio information. As pointed out by [39], the "audible but not visible events" are more common than "visible but not audible events" since the camera only captures the scene of limited view, while the microphones capture events from all directions. Thus, the visual modality is more likely to encounter the non-correlated cross-modal context. To alleviate such situations, we introduce CAPC. Our idea is to allow the visual modality to learn from beyond its paired audio, and induce it to have consistent visual event predictions as learning with its original pair. As such, the visual stream learns to only fuse the audio context that is correlated with itself and ignores other unrelated information to maintain the same visual event detection under different audio contexts.

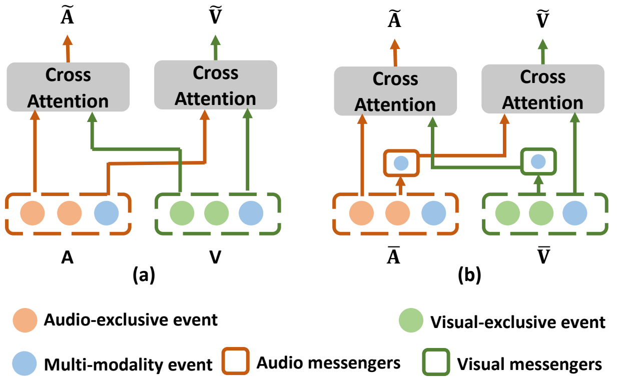

Figure 2. Network comparison: (a) HAN embedding and (b) our embedding. A and V denote the pre-extracted audio and visual features, respectively. $\overline{A}$ and $\overline{V}$ denotes the self-refined audio and visual features, respectively. $\tilde{A}$ and $\tilde{V}$ denote the audio and visual representation after cross-modal fusion, respectively. Audio-exclusive (visual-exclusive) events denotes events only happening on audio (visual) modality. Each circle represents one event.

In summary, our contributions are as follows:

- We propose messenger-guided mid-fusion transformer to reduce the uncorrelated cross-modal context in the audio-visual fusion. The messengers serve as a compact representation to amplify the most relevant cross-modal information under weak supervision.
- We propose cross-audio prediction consistency to calibrate the visual event prediction under interference from the unmatched audio context. The visual event prediction is forced to remain unchanged when pairing with different audios so as to ignore the irrelevant audio information.
- We conduct extensive qualitative and quantitative experiments to analyze the effectiveness of our approach. Our method also achieves the state-of-the-art performance on AVVP benchmark.

2. Related Work

2.1. Audio-Visual Representation Learning

The natural correspondence between audio and visual modalities overcomes the limitations of perception tasks in single modality and introduces a series of new applications. Semantic similarity is the most commonly used audio-visual correlation [2,3,5,20,28]. The shared semantic information in both audio and visual modality is a valuable free-source supervision. SoundNet [5] learns sound

[1] We change the input of the cross-attention module to the modality itself so as to keep the model size unchanged.

representation by distilling knowledge from the pre-trained visual models. Morgado *et al*. [28] propose cross-modal contrastive learning, where negative and positive samples of visual frames are drawn from audio samples and vice-versa. Besides semantic correlation, other works utilize temporal synchronization [1,22,30,32], motion correlation [12,41] and spatial correspondence [13,26,40]. However, self-supervised learning from natural videos is potentially noisy as the audio and visual modalities are not always correlated. Recently, Morgado *et al*. [27] propose to learn robust audio-visual representation by correcting the false alignment in the contrastive loss. Our task shares similar motivation with [27] as some events only happen in single modality leaving no audio-visual correspondence.

2.2. Audio-Visual Video Parsing

Early works [23,35] only detect events that is both audible and visible. Based on the strong assumption that the audio and visual information are aligned at each time step, [23, 35] fuse the audio and visual features at the same time step. However, the events happening on the two modalities are not always the same since the audio and vision are inherently different sensors. To fully understand the content in the multimodal videos, Tian *et al*. [34] introduced the task of audio-visual video parsing (AVVP). It classifies and localizes all the events happening on the audio and visual streams in a weakly-supervised manner. They design a hybrid attention network (HAN) to capture the uni-modal and cross-modal temporal contexts simultaneously. The audio and visual features are fused at the start of the network, where the self-attention and cross-attention are performed in parallel. Since then, HAN serves as the state-of-the-art audio-visual embedding and is widely adopted in follow-up works. MA [39] generates reliable event labels for each modality by exchanging the audio and visual tracks of a training video with another unrelated video. JoMoLD [8] leverage audio and visual loss patterns to remove modality-specific noisy labels for each modality. Lin *et al*. [24] explore the cross-modality co-occurrence and shared cross-modality semantics across-videos. Although HAN embedding shows promising performance, the full entanglement of two non-fully correlated modalities is not ideal for the task of AVVP. Therefore, we propose messenger-guided mid-fusion transformer and cross-audio prediction consistency to reduce the uncorrelated cross-modal context in the audio-visual fusion.

2.3. Multimodal Transformer

The attention module in the transformer [37] is effective in capturing the global context among the input tokens, and is widely adopted in the multimodal task [11, 16, 21, 25, 29]. Gabeur *et al*. [11] use transformer to capture cross-modal cues and temporal information in the video. OMNIVORE [16] proposes a modality-agnostic visual model that can perform classification on image, video, and single-view 3D modalities using the same shared model parameters. Perceiver [21] and MBT [29] address the high computation cost of the multimodal transformer by using a small set of fusion tokens as the attention bottleneck to iteratively distill the uni-modal inputs. Despite our messengers also serve as the attention bottleneck, it is used to suppress learning from noisy labels. Moreover, the messengers are more effective in the small multimodal models. The MBT and Perceiver initialize the fusion tokens randomly and require multiple times updates with the uni-modal inputs for it to carry meaningful cross-modal information, which is not applicable for the model with small number of encoder layers. In contrast, our messenger is directly derived from the global representation of each modality so that it is already informative without multiple times of updates.

3. Method

Let us denote a video with T non-overlapping segments as $\{\mathrm{V}_t, \mathrm{A}_t\}_{t=1}^{T}$, where V_t and A_t are the visual and audio clip at the t-th segment, respectively. The corresponding label for visual event, audio event and audio-visual event at the t-th segment is denoted as $y_t^{\mathrm{v}} \in \{0,1\}^C$, $y_t^{\mathrm{a}} \in \{0,1\}^C$ and $y_t^{\mathrm{av}} \in \{0,1\}^C$, respectively. C is the total number of classes in the dataset and $y_t^{\mathrm{av}} = y_t^{\mathrm{v}} \times y_t^{\mathrm{a}}$. An event is considered as an audio-visual event only if it occurs in both modalities. Note that more than one event can happen in each segment. Grouping all the segment-level labels together, we obtain the video-level label $Y = \{y_t^{\mathrm{v}} \cup y_t^{\mathrm{a}}\}_{t=1}^{T} \in \{0,1\}^C$. The goal of audio-visual video parsing is to detect all the visual, audio and audio-visual events in the video. The training of AVVP is conducted in weak supervision, where only video-level labels Y are provided.

In the following sections, we first introduce messenger-guided mid-fusion transformer as the new multi-modal embedding for AVVP, and then the novel idea of cross-audio prediction consistency to reduce the negative interference of the unmatched audio context to the visual stream.

3.1. Messenger-guided Mid-Fusion Transformer

Fig. 3 shows our proposed messenger-guided mid-fusion transformer. We instantiate the self-attention and cross-attention layers with transformers [37] as it shows excellent performance in the uni-modal [4, 9] and multimodal [11, 16, 31] tasks with just the attention mechanism. The pre-trained visual and audio feature extractors extract segment-level visual features $\{f_t^{\mathrm{v}}\}_{t=1}^{T}$ and audio features $\{f_t^{\mathrm{a}}\}_{t=1}^{T}$, respectively. The $\{f_t^{\mathrm{v}}\}_{t=1}^{T}$ and $\{f_t^{\mathrm{a}}\}_{t=1}^{T}$ are the input to the multimodal embedding, where the uni-modal and cross-modal context are modeled sequentially. Instead of directly feeding the full cross-modal context to the fusion, we summarize it into compact messengers. Finally, the outputs of

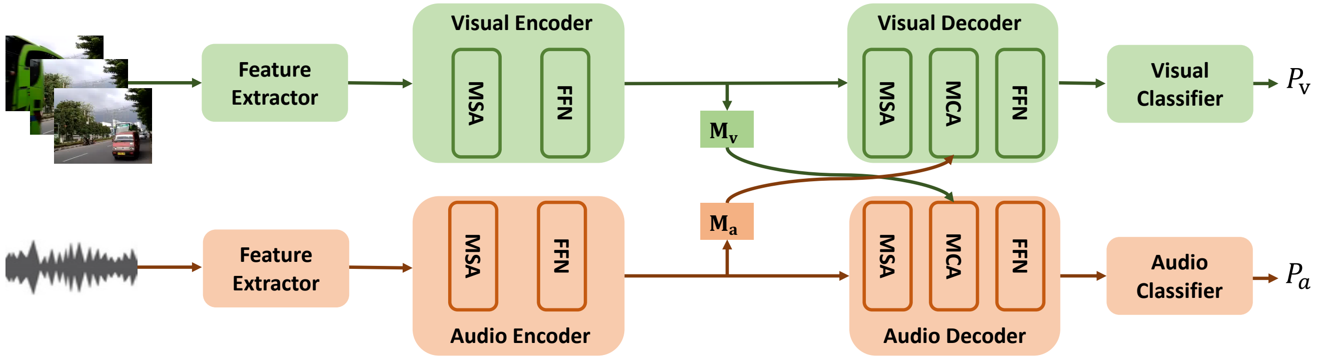

Figure 3. The architecture of the messenger-guided mid-fusion transformer. Visual Encoder and Audio Encoder denote the transformer encoders for the visual and audio, respectively. Visual Decoder and Audio Decoder denote the transformer decoders for the visual and audio, respectively.

the last layer of the decoders are sent into the classifiers to detect segment-level events for each modality.

Uni-modal Context Refinement. We model the uni-modal context with transformer encoders, where the self-attention is capable of aggregating the temporal context for a better global understanding of the raw input sequence. For brevity, we only illustrate the working flow of the visual modality since the visual and audio branches work symmetrically. The pre-extracted visual features are first converted into 1-D tokens $S_{\text{v}} \in \mathbb{R}^{T \times d}$ with feature dimension d as follow:

$$S_{\text{v}} = [f_1^{\text{v}}\mathbf{W}_{\text{v}}^{\text{enc}}, f_2^{\text{v}}\mathbf{W}_{\text{v}}^{\text{enc}}, \dots, f_{\text{T}}^{\text{v}}\mathbf{W}_{\text{v}}^{\text{enc}}] + \mathbf{PE}, \tag{1}$$

where $\mathbf{W}_{\text{v}}^{\text{enc}}$ is the linear projection layer that projects pre-extracted features to d dimension and $\mathbf{PE}$ is the position embedding. Then, the tokens are sent into a L-layer transformer encoder. We adopt the original architecture of transformer [37]. Each layer consists of a multi-headed self-attention (MSA) and a position-wise fully connected feed-forward network (FFN):

$$\begin{aligned} \tilde{S}_{\text{v}}^{l} &= \text{LN}\left(\text{MSA}\left(S_{\text{v}}^{l}\right) + S_{\text{v}}^{l}\right), \\ S_{\text{v}}^{l+1} &= \text{LN}\left(\text{FFN}\left(\tilde{S}_{\text{v}}^{l}\right) + \tilde{S}_{\text{v}}^{l}\right), \end{aligned} \tag{2}$$

where LN denotes layer normalization and S_{v}^{l} is the input tokens at the l-th layer.

Cross-modal Context Fusion with Messengers. We model the cross-modal context using M-layer transformer decoders, where each layer consists of the multi-headed self-attention (MSA), multi-headed cross-attention (MCA), and the position-wise feed-forward network (FFN). The working flow inside the m-th layer of the decoder is as follows:

$$\begin{aligned} \tilde{R}_{\text{v}}^{m} &= \text{LN}\left(\text{MSA}\left(R_{\text{v}}^{m}\right) + R_{\text{v}}^{m}\right), \\ \hat{R}_{\text{v}}^{m} &= \text{LN}\left(\text{MCA}\left(\tilde{R}_{\text{v}}^{m}, S_{\text{a}}^{L}\right) + \tilde{R}_{\text{v}}^{m}\right), \\ R_{\text{v}}^{m+1} &= \text{LN}\left(\text{FFN}\left(\hat{R}_{\text{v}}^{m}\right) + \hat{R}_{\text{v}}^{m}\right), \end{aligned} \tag{3}$$

where $\text{MCA}(\cdot)$ performs cross-modal fusion between the visual feature $\tilde{R}_{\text{v}}^{m}$ (query) and the audio context S_{a}^{L} (key and value).

However, providing full cross-modal context is not ideal when the two modalities are not fully correlated, and is even worse when the supervision is noisy. The video-level label Y is the union of the audio and visual events, which can introduce noise when supervising each modality, *i.e.* an event that is present only in one modality becomes a noisy label for the other modality. As shown in Fig. 4(b), with connection with full audio context, the model is assured that the audio-exclusive event truly happens in the visual stream guided by the noisy supervision Y. Consequently, its generalization ability is severely affected. To this end, we create a fusion bottleneck M_{a} as shown in Fig. 4(c) so that they can suppress the irrelevant audio context, and the model is less likely to overfit to the noisy label Y.

Specifically, we condense the full cross-modal context into the compact representation M_{v} as follows:

$$M_{\text{v}} = \text{Tanh}\left(\text{Pool}\left(S_{\text{v}}^{L}\mathbf{W}_{\text{v}}^{\text{msg}}; n_{\text{v}}\right)\right), \tag{4}$$

where $\mathbf{W}_{\text{v}}^{\text{msg}} \in \mathbb{R}^{d \times d}$ is the linear projection layer, $\text{Pool}(\cdot)$ is the average pooling along the temporal dimension with a target length of n_{v} and $\text{Tanh}(\cdot)$ is the activation function. $M_{\text{v}} \in \mathbb{R}^{n_v \times d}$ has limited capacity in storing information compared to the full cross-modal context $S_{v}^{L} \in \mathbb{R}^{T \times d}$, where T is much larger than n_v. Consequently, it creates an attention bottleneck that gives priority to the most relevant cross-modal context that fits the clean labels. The compact audio context $M_{\text{a}} \in \mathbb{R}^{n_a \times d}$ at the audio stream is obtained

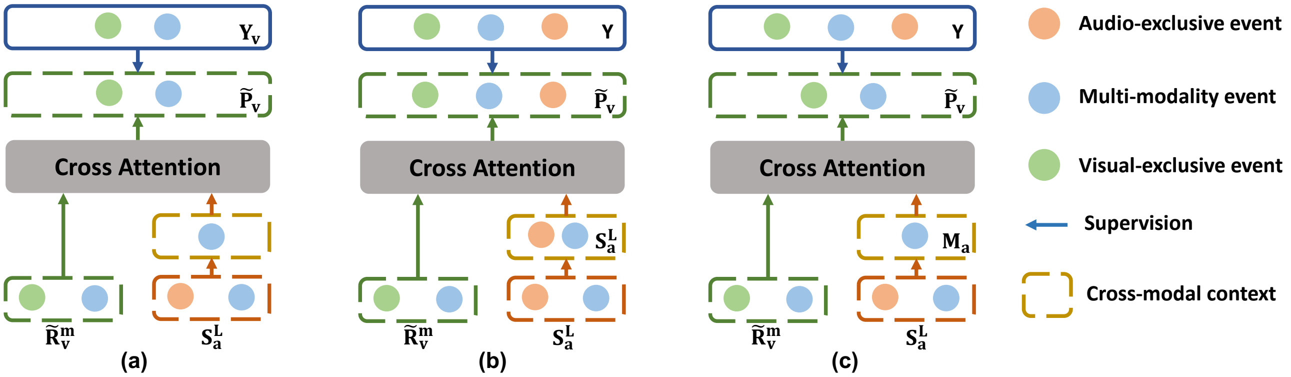

Figure 4. Comparison of different fusion methods. Each circle represents one event. We only illustrate the cross-modal fusion at the visual branch for brevity. (a) shows an oracle setting, where the visual label Y_{v} is available. In this case, the model only learns from the relevant audio context. (b) shows the fusion with full audio context. (c) shows the fusion with messengers.

in a similar way. We name it messenger as it represents its source modality as the direct input in the cross-modal fusion as follows:

$$\hat{R}_{\text{v}}^{m} = \text{MCA}\left(\text{LN}\left(\tilde{R}_{\text{v}}^{m}, M_{\text{a}}\right)\right) + \tilde{R}_{\text{v}}^{m}, \tag{5}$$

where the audio messengers M_{a} replace the full cross-modal context S_{a}^{L}.

Classification. R_{v}^{M} is considered as the final visual representation and is sent into the modality-specific classifiers for segment-level event prediction $P_{\text{v}} \in \mathbb{R}^{T\times C}$ as follow:

$$P_{\text{v}} = \text{Sigmoid}\left(R_{\text{v}}^{M}\mathbf{W}_{\text{v}}^{\text{cls}}\right), \tag{6}$$

where $\mathbf{W}_{\text{v}}^{\text{cls}} \in \mathbb{R}^{d\times c}$ is the classifier weights. During training, we aggregate P_{v} into video-level predictions $\tilde{P}_{\text{v}} \in \mathbb{R}^{C}$ via soft pooling similar to [34] since only video-level label $Y \in \{0,1\}^{C}$ is provided. The audio prediction P_{a} and $\tilde{P}_{\text{a}}$ are obtained in a similar way. We also combine $\tilde{P}_{\text{v}}$ and $\tilde{P}_{\text{a}}$ into a modal-agnostic video-level prediction $\tilde{P}_{\text{video}} \in \mathbb{R}^{C}$. In total, we have three classification losses:

$$\mathcal{L}_{\text{cls}} = \text{CE}\left(\tilde{P}_{\text{v}}, Y_{\text{v}}\right) + \text{CE}\left(\tilde{P}_{\text{a}}, Y_{\text{a}}\right) + \text{CE}\left(\tilde{P}_{\text{video}}, Y\right), \tag{7}$$

where CE denotes the binary cross-entropy loss. We set $Y_{\text{v}} = Y_{\text{a}} = Y$ when only the video-level label is available.

3.2. Cross-Audio Prediction Consistency

We further propose cross-audio prediction consistency (CAPC) to suppress the inaccurate visual event prediction arising from unmatched audio information. As analyzed in [39], audio-exclusive events are more common than visual-exclusive events and thus the visual stream is more likely to be influenced by the non-correlated cross-modal context. As shown in Fig. 6, the visual branch confidently detects the audio-exclusive event 'Chicken rooster' when only learnt with its paired audio. To alleviate this problem, we introduce a consistency loss in the visual event prediction by pairing the same visual sequence with different audios. Specifically, the visual modality V is paired with not only its original audio counterpart A_{orig}, but is also paired with audios that are randomly selected from other videos at each training iteration. We denote the visual prediction from the original pair $(\text{V}, \text{A}_{\text{orig}})$ as $\tilde{P}_{\text{v}} \in \mathbb{R}^{C}$, and the visual prediction from the i-th random pair $(\text{V}, \text{A}_{\text{rand}}^{i})$ as $\tilde{P}_{\text{v}}^{i} \in \mathbb{R}^{C}$. CAPC requires $\tilde{P}_{\text{v}}^{i}$ to be the same as $\tilde{P}_{\text{v}}$ as follow:

$$\mathcal{L}_{\text{ccr}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{P}_{\text{v}}^{i} - \tilde{P}_{\text{v}}\right\|_{2}^{2}, \tag{8}$$

where N is the number of random pairs for each visual sequence. The cross-attention at the visual stream will learn to only grab the useful audio context (*i.e.* audio-visual event) from A and $\text{A}_{\text{rand}}^{i}$ and ignore the irrelevant information (*i.e.* audio-exclusive event) in order to achieve this prediction consistency.

We notice that there may be a trivial solution to achieve this prediction consistency. The cross-attention totally ignores all the audio context and thus leads to complete independence of the visual prediction from the audio information. However, we show in Tab. 4 that the model does not degenerate to this trivial solution. Instead, CAPC improves the robustness of fusion under non-fully correlated cross-modal context.

Finally, the total loss of our method is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mu\mathcal{L}_{\text{ccr}}, \tag{9}$$

where μ is the hyperparameter to balance the loss terms.

Method	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event
AVE [35]	47.2	40.4	37.1	34.7	35.4	31.6	39.9	35.5	41.6	36.5
AVSDN [23]	47.8	34.1	52.0	46.3	37.1	26.5	45.7	35.6	50.8	37.7
HAN [34]	60.1	51.3	52.9	48.9	48.9	43.0	54.0	47.7	55.4	48.0
HAN † [34]	59.8	52.1	57.5	54.4	52.6	45.8	56.6	50.8	56.6	49.4
MA [39]	60.3	53.6	60.0	56.4	55.1	49.0	58.9	53.0	57.9	50.6
Lin *et al.* [24]	60.8	53.8	63.5	58.9	57.0	49.5	60.5	54.0	59.5	52.1
JoMoLD [8]	61.3	**53.9**	63.8	59.9	57.2	49.6	60.8	54.5	59.9	52.5
Ours	**61.9**	**53.9**	**64.8**	**61.6**	**57.6**	**50.2**	**61.4**	**55.2**	**60.9**	**53.1**

Table 2. Comparison with the state-of-the-art methods of audio-visual video parsing on the LLP test dataset. 'Audio', 'Visual' and 'Audio-Visual' denotes audio event, visual event and audio-visual event detection, respectively. Note that they are different from the event categories in Tab. 1 and we illustrate the difference in the Supplementary Material. 'Seg.' denotes segment-level evaluation and 'Event' denotes event-level evaluation. 'HAN†' is the variant of HAN that additionally uses label refinement. The best result is marked in bold.

3.3. Discussions on CAPC

Comparison with Consistency Regularization. Consistency regularization is widely adopted in semi-supervised learning [6, 7, 33], where the model is required to output the similar prediction when fed perturbed versions of the same image. In contrast to augmenting the modality itself, the CAPC keeps the modality itself intact and only augments its cross-modal context, *i.e.* pairing the visual modality with different audios. By learning consistency under cross-audio augmentation, the fusion robustness is improved.

Comparison with Audio-Visual Correspondence. Audio-visual pairing correspondence [2, 3] and audio-visual temporal correspondence [22, 30] are the cross-modal self-supervision in the video. They highlight the audio-visual alignment to learn good audio and visual representation. In contrast, we highlight the audio-visual misalignment to reduce the negative impact of the unmatched audio context to the visual modality.

4. Experiments

4.1. Experiment Setup

Dataset. We conduct experiments on the *Look, Listen and Parse (LLP) Dataset* [34]. It contains 11,849 YouTube videos, each is 10-second long. It covers 25 real-life event categories, including human activities, animal activities, music performances, *etc*. 7,202 video clips are labeled with more than one event category and per video has an average of 1.64 different event categories. We follow the official data splits [34]. 10,000 video clips only have video-level labels and are used as training sets. The remaining 1,849 videos that are annotated with audio and visual events and secondwise temporal boundaries are divided into 849 videos as the validation set and 1,000 as the test set.

Evaluation Metrics. We use F-scores as the quantitative evaluation method. We parse the visual, audio and audio-visual events, denoted as 'Visual', 'Audio' and 'Audio-Visual', respectively. We use F-scores for segment-level and event-level evaluations. The segment-level metrics evaluate the predictions for each segment independently. The event-level evaluation first concatenates consecutive positive segments as the event proposal and then compares the alignment with the ground-truth event snippet under the mIoU=0.5 as the threshold. Meanwhile, we use 'Type@AV' and 'Event@AV' for the overall access of the model performance. The 'Type@AV' averages the evaluation scores of 'Audio', 'Visual' and 'Audio-Visual. The 'Event@AV' considers all the audio and visual events for each video rather than directly averaging results from different event types.

Implementation Details. Each video is downsampled at 8 fps and divided into 1-second segments. We use both the ResNet-152 [18] model pre-trained on ImageNet and 18 layer deep R(2+1)D [36] model pre-trained on Kinetics-400 to extract visual features. The 2D and 3D features are concatenated as the visual representation for the visual input. For the audio signals, we use the VGGish network [19] pre-trained on AudioSet [15] to extract 128-D features. The feature extractors are fixed during training. For each modality, both the number of encoders L and the number of decoders M are set to 1. The hidden size is set to 512 and the number of attention head is set to 1. The number n_a of audio messengers and the number n_v of visual messengers are set to 1, respectively.

We train our model in three stages. In the first stage, we optimize our proposed audio-visual embedding with classification loss $\mathcal{L}_{\text{cls}}$ on the video-level label Y. In the second stage, we calculate the pseudo label [39] for each modality. In the third stage, we re-train our embedding with Eqn. 9 using the pseudo label. μ is set to 0.5 and N is set to 1. We use Adam optimizer with learning rate 3×10^{-4} and batch size of 64. We train 40 epochs and decrease the learning rate by 10^{-1} every 10 epochs in each training stage. All the experiments are conducted using Pytorch on a NVIDIA GTX 1080 Ti GPU.

n_a	n_v	Audio	Visual	Audio-Visual	Type@AV	Event@AV
No MSG		61.5	63.2	55.7	60.1	59.9
MBT		60.9	64.1	56.0	60.3	59.8
5	5	61.5	64.4	57.0	61.0	60.6
3	3	**62.1**	63.9	56.9	61.0	60.5
1	3	61.6	64.7	56.8	61.0	60.6
3	1	**62.1**	64.6	56.8	61.2	**60.9**
1	1	61.9	**64.8**	**57.6**	**61.4**	**60.9**

Table 3. Ablation study of the messengers. 'No MSG' denotes the model performs cross-modal fusion without messengers. 'MBT' replaces the messengers with the fusion bottleneck token [29]. n_a and n_v is the number of audio and visual messengers, respectively, and '$n_\text{a} = 1, n_\text{v} = 1$' is our final model. Segment-level results are reported.

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

We compare our method with state-of-the-art audio-visual event parsing methods of AVE [35], AVSDN [23], HAN [34], MA [39], Lin *et al*. [24] and JoMoLD [8]. We report their results from their paper. All the methods are trained using the same pre-extracted features as input. The recent methods, MA [39], Lin *et al*. [24] and JoMoLD [8] all adopt the HAN [34] as the audio-visual embedding.

Tab. 2 shows the quantitative comparisons on the LLP dataset [34]. Our model constantly outperforms other methods on all the evaluation metrics. Although we remove the entanglement in the early layers, the performance of audio and visual event detection are both improved than all the HAN-based methods. This demonstrates that a compact fusion is better than the fully entangled fusion approach when the audio and visual information are not always correlated.

4.3. Ablation Study

Effectiveness of the Messengers. Tab. 3 shows the effectiveness of the messengers. 'No Messenger' abbreviated as 'No MSG' is the model which directly performs cross-modal fusion using Eqn. 3. Compared with our final model '$n_\text{a} = 1, n_\text{v} = 1$', both the audio and visual performance decrease. We also provide qualitative analysis of the messenger in Fig. 5. 'No MSG' wrongly detects the visible but not audible event 'Car' on the audio stream due to the unconstrained visual context. By constraining the full visual context into our compact messenger, the irrelevant visual information 'Car' is suppressed and only keeps the useful information of the 'Motorcycle' for the audio.

In addition, we compare our messenger with fusion bottleneck token [29], denoted as 'MBT', by replacing the messenger with the same number of fusion bottleneck tokens in the transformer model. Our messenger '$n_\text{a} = 1, n_\text{v} = 1$' consistently outperforms 'MBT' in the shallow transformer model, where only one layer of the encoder is used for each modality.

Analysis of the number of messengers. Tab. 3 also shows the performance of our model with different numbers of messengers. n_a and n_v is the number of audio and visual messengers, respectively. Using a large number of messengers shows decreasing performance, suggesting the dilution of the beneficial cross-modal context. Hence we adopt '$n_\text{a} = 1, n_\text{v} = 1$' in our final model.

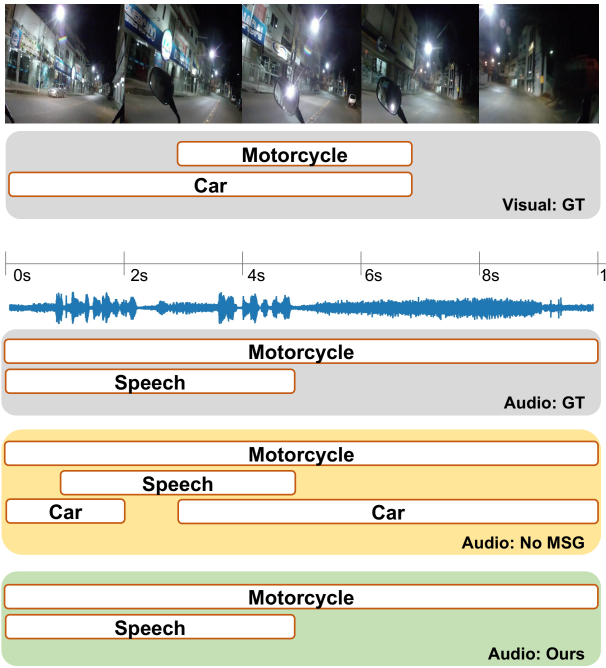

Figure 5. Qualitative comparison of the messengers. 'Visual' and 'Audio' represent the visual and audio event, respectively. 'GT' denotes ground truth. 'No MSG' denotes model performs cross-modal fusion without messengers. 'Ours' denotes the fusion with messengers.

	Audio	Visual	Audio-Visual	Type@AV	Event@AV
No FA	**62.1**	62.5	56.3	60.3	60.3
No CAPC	61.9	64.2	56.4	60.8	60.7
Ours	61.9	**64.8**	**57.6**	**61.4**	**60.9**

Table 4. Ablation study of the cross-audio prediction consistency. 'No FA' denotes the model without fusion with audio at the visual stream. 'No CAPC' denotes the model without using cross-audio prediction consistency. Segment-level results are reported

Effectiveness of Cross-audio Prediction Consistency. Tab. 4 presents the ablation study of cross-audio prediction consistency. 'No CAPC' is the model trained without cross-audio prediction consistency. By forcing prediction consistency on the visual stream, *i.e*. 'Ours', both the performance of visual and audio-visual event detection show obvious improvement.

We also verify whether CAPC leads to the trivial solution, *i.e*. the visual prediction does not need the audio information at all. We replace the input to the cross-attention at the visual

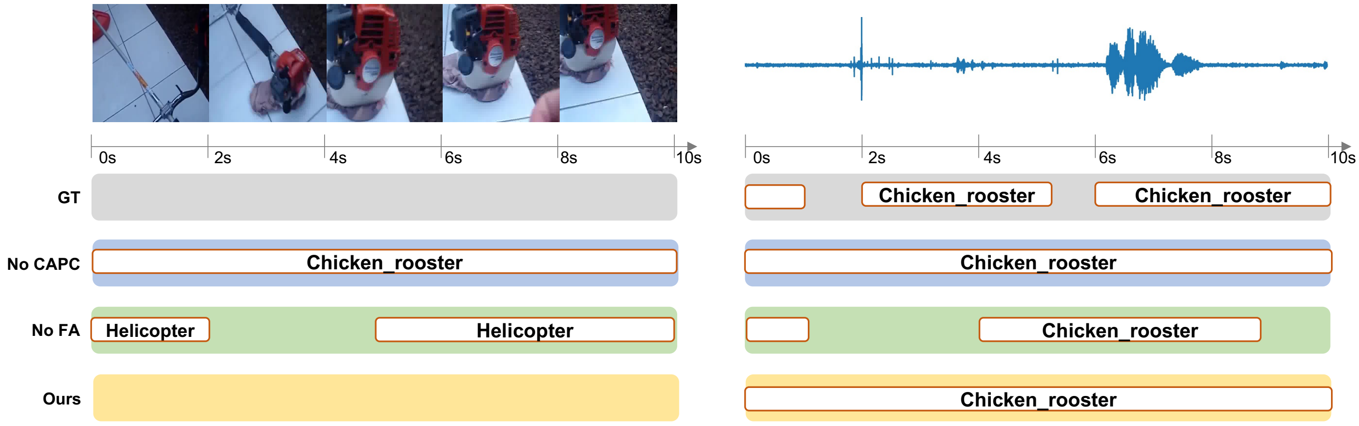

Figure 6. Qualitative comparison of CAPC. 'GT' denotes the ground truth event labels. 'No CAPC' denotes the model without cross-audio prediction consistency. 'No FA' denotes the model without fusion with audio at the visual stream.

μ	Audio	Visual	Audio-Visual	Type@AV	Event@AV
0.1	61.3	63.6	56.3	60.4	60.1
0.5	**61.9**	**64.8**	**57.6**	**61.4**	**60.9**
1	**61.9**	64.6	**57.6**	**61.4**	60.7

Table 5. Analysis of different value of μ. Segment-level results are reported.

N	Audio	Visual	Audio-Visual	Type@AV	Event@AV
1	**61.9**	**64.8**	**57.6**	**61.4**	**60.9**
2	61.7	64.5	57.0	61.1	60.5
3	61.7	64.0	57.0	60.9	60.6

Table 6. Analysis of different number of random pairs N. Segment-level results are reported.

stream with the visual modality itself, *i.e.* replacing M_{a} with S_{v}^L in Eqn. 5, denoted as 'No FA'. The performance is worse than the 'No CAPC', and much worse than our full model 'Ours'. We also provide the qualitative comparison in Fig. 6. 'No FA' shows the poor generalization ability on the visual event detection as it wrongly detects a totally irrelevant event 'Helicopter'. By learning with both its paired audio and other non-paired audios, the model can correctly identify that the 'Chicken rooster' is an audible but not a visible event. This shows that our cross-audio prediction consistency does not de-activate the cross-modal fusion. Instead, it improves the robustness of the fusion between two non-fully correlated modalities.

Analysis of CAPC loss weight μ. Tab. 5 presents the ablation study on different value of μ. Interestingly, using a small weight, *i.e.* $\mu = 0.1$ is worse than the model without CAPC ('No CAPC' in Tab. 4). The possible reason is that the model trained with '$\mu = 0.1$' tends to focus on the easy training samples (A$_{\text{orig}}$ and A$_{\text{rand}}^i$ are both fully correlated with V). CAPC thus encourages V to take in full audio context to achieve faster convergence in this case, which provides the false signal. Only using larger weight can effectively optimize hard samples (A$_{\text{orig}}$ and A$_{\text{rand}}^i$ are not fully correlated with V), where CAPC guides visual stream to only select useful audio information. It can also be verified that CAPC loss of $\mu = 0.1$ is much larger than $\mu = 0.5$. Therefore, we choose $\mu = 0.5$ as our final setting.

Analysis of the number of random pairs in CAPC. Tab. 6 presents the ablation study of the number of random pairs in CAPC. Using a larger number of pairs, *i.e.*, $N = 3$, not only increases computation cost but also exhibits a decline in performance. We postulate the reason is that larger N provides a false signal that the visual modality should ignore any audio context (including correlated audio information). More analysis is provided in Supplementary Material. Therefore, we choose $N = 1$ in our final model.

5. Conclusion

We address the problem of fusion between two non-fully correlated modalities in weakly supervised audio-visual video parsing. We propose the messenger-guided mid-fusion transformer to reduce the unnecessary cross-modal entanglement. The messengers act as fusion bottlenecks to mitigate the detrimental effect of the noisy labels. Further, we propose cross-audio prediction consistency to reduce the negative interference of unmatched audio context to the visual stream. The effectiveness of our proposed method is analyzed both quantitatively and qualitatively.

Acknowledgement. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-024), and the Tier 2 grant MOE-T2EP20120-0011 from the Singapore Ministry of Education.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 3

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 2, 6

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2, 6

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 3

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 2

[6] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 6

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 6

[8] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 431–448. Springer, 2022. 3, 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013. 1

[11] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 1, 3

[12] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 3

[13] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020. 3

[14] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[16] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022. 3

[17] Wangli Hao, Zhaoxiang Zhang, He Guan, and Guibo Zhu. Integrating both visual and audio cues for enhanced video caption. In *AAAI Conference on Artificial Intelligence*, 2018. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[20] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[21] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021. 3

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 3, 6

[23] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 3, 6, 7

[24] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 6, 7

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[26] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 3

[27] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021. 3

[28] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2, 3

[29] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 7

[30] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 3, 6

[31] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 3

[32] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. Multimodal video summarization via time-aware transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1756–1765, 2021. 3

[33] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 6

[34] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 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*, pages 436–454. Springer, 2020. 1, 2, 3, 5, 6, 7

[35] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1, 3, 6, 7

[36] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4

[38] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020. 1

[39] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 2, 3, 5, 6, 7

[40] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 3

[41] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 3