Cross-Modal Label Contrastive Learning for Unsupervised Audio-Visual Event Localization

Peijun Bao[1], **Wenhan Yang**[*1,2], **Boon Poh Ng**[1], **Meng Hwa Er**[1], **Alex C. Kot**[1]
[1]Nanyang Technological University
[2]Peng Cheng Laboratory
peijun001@e.ntu.edu.sg, yangwh@pcl.ac.cn, {ebpng, emher, eackot}@ntu.edu.sg

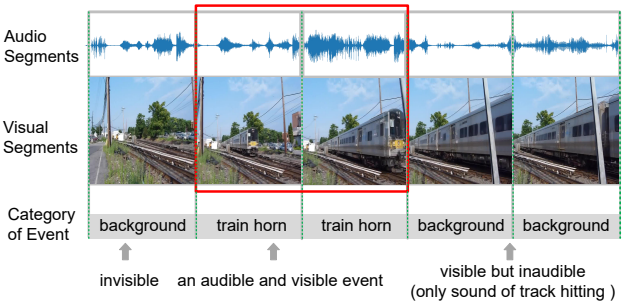
Abstract

This paper for the first time explores audio-visual event localization in an unsupervised manner. Previous methods tackle this problem in a supervised setting and require segment-level or video-level event category ground-truth to train the model. However, building large-scale multi-modality datasets with category annotations is human-intensive and thus not scalable to real-world applications. To this end, we propose cross-modal label contrastive learning to exploit multi-modal information among unlabeled audio and visual streams as self-supervision signals. At the feature representation level, multi-modal representations are collaboratively learned from audio and visual components by using self-supervised representation learning. At the label level, we propose a novel self-supervised pretext task i.e. label contrasting to self-annotate videos with pseudo-labels for localization model training. Note that irrelevant background would hinder the acquisition of high-quality pseudo-labels and thus lead to an inferior localization model. To address this issue, we then propose an expectation-maximization algorithm that optimizes the pseudo-label acquisition and localization model in a coarse-to-fine manner. Extensive experiments demonstrate that our unsupervised approach performs reasonably well compared to the state-of-the-art supervised methods.

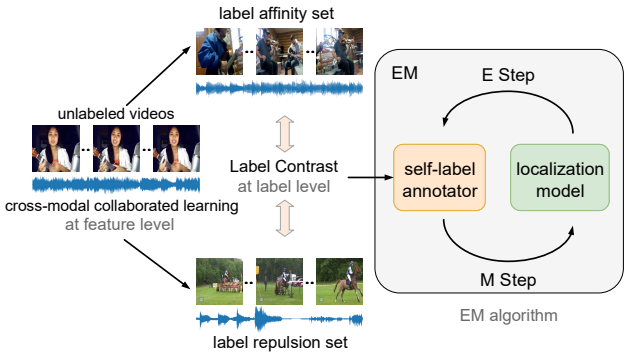
Introduction

Over the last few years, the computer vision community has witnessed the success of audio-visual event localization (Wang et al. 2021; Tian et al. 2018; Lin, Li, and Wang 2019; Xuan et al. 2020; Xu et al. 2020; Zhou et al. 2021; Ramaswamy 2020; Ramaswamy and Das 2020). In this task, an audio-visual event refers as an event that is both visible and audible in a video. And as illustrated in Fig 1 (a), the goal is to find which temporal segment contains an audio-visual event and identify what category the event belongs to.

Existing works tackle the audio-visual localization in a supervised manner (Wang et al. 2021; Tian et al. 2018; Lin, Li, and Wang 2019; Xuan et al. 2020; Xu et al. 2020; Zhou et al. 2021), and the manual annotations of segment-level or video-level event categories are required to train the model.



(a) An example of audio-visual event localization



(b) The pipeline of cross-modal label contrastive learning with expectation-maximization

Figure 1: We for the first time propose to solve audio-visual event localization in the unsupervised setting. (a) An example of audio-visual event localization, whose model training relies on heavy annotations. (b) The proposed pipeline for unsupervised audio-visual event localization, with self-supervision at both feature/label level, and an EM algorithm.

However, collecting a large number of videos with associated ground-truth labels is often time-consuming and labor-intensive, which makes it not scalable to real-world applications. On the contrary, it costs much less human effort to collect videos in multi-modalities, i.e. RGB frames with audio signals, which can often be easily obtained on web resources. Motivated by this, we pose a question in this paper: is it possible to develop an unsupervised framework for audio-visual event localization by leveraging the visual and

*Corresponding author.

audio component of videos without associated event category labels?

Our observation is that humans are not only able to perceive the world through simultaneous sensory streams but can further learn from the multi-modal information (Bulkin and Groh 2006; Smith and Gasser 2005). This inspires us to solve the problem of audio-visual event localization by exploiting collaborative supervision signals from the unlabeled audio and visual streams, without using video-level or segment-level annotations. To the best of our knowledge, we are the first in the literature to address this problem in an unsupervised setting.

Specifically, as illustrated in Fig 1 (b), we devise a cross-modal label contrastive learning framework with expectation maximization. At the feature representation level, the cross-modal collaborative representations are first learned from the unlabeled audio and visual streams by utilizing two pretext tasks i.e. instance discrimination and feature decorrelation. At the label level, we propose a novel self-supervised label contrastive learning to automatically discover event classes from the multi-modal features. Different from the conventional contrastive learning to learn feature representations, our label contrastive learning aims to contrast event category label distributions and learn the reliable pseudo-labels with self-supervision. Instead of obtaining positive / negative pairs with data augmentation, we first require to construct contrastive pairs by mining the label affinity and repulsion set from the whole training dataset. Then a label contrasting loss trains a self-label annotator to narrow the gap of pseudo-label distributions among affinity sets, and distance it in the repulsion set vice versa. The localization model training can then be formulated as multiple instance learning (Maron and Lozano-Pérez 1998) with the generated video level pseudo-labels.

Note that the training videos consist of both foreground events and background segments as shown in Fig 1 (a). These irrelevant backgrounds introduce noise to both audio and visual features, which hinders the acquisition of precise pseudo-labels. To remove these background segments, one requires to train a powerful localization model, which in turn relies on the high quality of pseudo labels. To address this issue, we propose an elegant Expectation-Maximization (EM) algorithm which regards the pseudo-labels as latent variables. In the expectation step (E-step), the pseudo-label distributions are evaluated from the current parameters of the localization model. And then the maximization step (M-step) re-estimates the parameters of localization model with the pseudo-labels from the E-step. The E-step and M-step iteratively run until the convergence of the localization model.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to explore unsupervised audio-visual event localization in the literature.
- We propose a novel self-supervised pretext task of label contrasting to mine event classes from cross-modal collaborative features, with label affinity and repulsion mechanism.
- An elegant EM algorithm treating pseudo-labels as latent variables is proposed to address the irrelevant background problem in a coarse-to-fine manner.
- The experiments show our unsupervised approach achieves comparative performance to the state-of-the-art supervised methods on the standard large-scale dataset of audio-visual event localization.

Related Works

Audio-visual event localization. Audio-visual event localization (Wang et al. 2021; Lin, Li, and Wang 2019; Xuan et al. 2020; Xu et al. 2020; Zhou et al. 2021) aims to predict which temporal segment of a video input has an audio-visual event and classify the category of the event. Existing works focus on the fully / weakly- supervised settings. Tian et al. (Tian et al. 2018) first tackle this task with a dual multimodal residual network to fuse the audio-visual information for localization in a supervised manner. Some later works explore the fusion strategy for multi-modalities. Xuan et al. (Xuan et al. 2020) propose a cross-modal attention network to adaptively trade information between audio and visual components. Similarly, a cross-modal relation-aware network is devised in (Xu et al. 2020) to enable simultaneous reasoning between the visual and audio content. A positive sample propagation module is proposed in (Zhou et al. 2021) to discover similar audio-visual pairs for better feature representation learning. Wu et al. (Wu et al. 2019) propose a dual attention matching module to perceive both long-term video contents and local temporal information for a better semantic understanding of audio-visual event.

Unsupervised audio-visual learning. The intrinsic correspondence between sound and vision provides effective supervision signals. A list of tasks ranging from sound object localization, sound source separation to object segmentation and sound event detection are tackled with self-supervised or unsupervised audio-visual learning (Hu, Nie, and Li 2019; Cheng et al. 2020; Rouditchenko et al. 2019; Alwassel et al. 2020). Hu et al. (Hu, Nie, and Li 2019) propose a deep multimodal clustering method for capturing multiple audio-visual correspondences and then apply it to sound source localization and sound event detection. Despite that the above tasks are already handled in unsupervised setting, temporal localization of audio-visual events in videos is still an unexplored one, and we are the first to tackle this task in the unsupervised case.

Unsupervised action localization. The most related works in visual action localization are (Soomro and Shah 2017; Gong et al. 2020). Soomro et al. (Soomro and Shah 2017) propose an unsupervised spatio-temporal action detection model without using bounding box and action categories. Actions are firstly grouped into identical categories and then these actions are utilized to train the model to detect video tubes that contain the actors performing the actions. Similarly, Gong et al. (Gong et al. 2020) devise a temporal co-attention model for unsupervised action localization. We emphasize that our works are fundamentally different from theirs in that we collaboratively exploit two modalities as self-supervision signals, while theirs focus on a single visual modality.

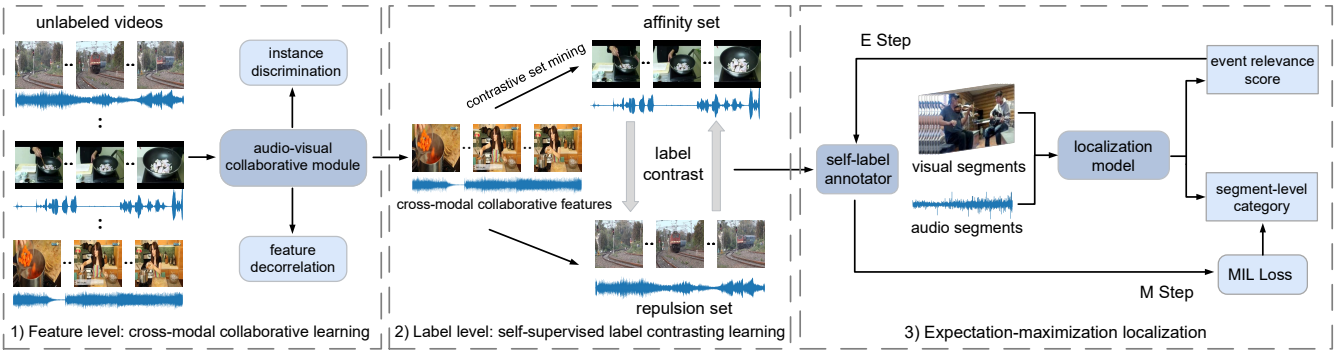


Figure 2: The framework of cross-modal label contrastive learning for unsupervised audio-visual event localization. 1) At feature representation level, the cross-modal collaborative learning aims to learn representations with two pretext tasks i.e. instance discrimination and feature decorrelation. 2) At label level, a novel label contrasting pretext task trains the model to generate pseudo-labels with contrasting between affinity and repulsion set. 3) To address the irrelevant background problem, an EM algorithm optimizes the pseudo-label acquisition and localization model in a coarse-to-fine manner.

Problem Description

The task of audio-visual event localization aims to predict which temporal segment of a video input has an audio-visual event and identify what category the event belongs to. Specifically, a video sequence $\mathcal{S}$ is splitted as T non-overlapping segment $\{S_t = (V_t, A_t)\}_{t=1}^{T}$, where V_t, A_t are the visual and audio components, respectively. The details of these three settings of supervision are described as the following.

Fully-supervised setting. The segment-level event labels can be accessed for training in the fully-supervised setting. In more specific, each segment-level event label is denoted as $\boldsymbol{y}_t = \{y_t^c \mid y_t^c \in \{0,1\}, \sum_{c=1}^{C} y_t^c = 1\} \in \mathbb{R}^C$, where C is the total number of event categories plus one background class. Then the label for the entire video can be given as $\boldsymbol{Y}^{\text{fully}} = [\boldsymbol{y}_1; \boldsymbol{y}_2; \ldots; \boldsymbol{y}_T] \in \mathbb{R}^{T \times C}$.

Weakly-supervised setting. In the weakly-supervised setting, the model can only access video-level ground truth $\boldsymbol{Y}^{\text{weak}} = \{y^c \mid y^c \in \{0,1\}, \sum_{c=1}^{C} y^c = 1\} \in \mathbb{R}^C$, instead of the segment-level one.

Unsupervised setting. Our paper explores a new unsupervised setting for audio-visual event localization. In the unsupervised case, we access neither the segment-level nor video-level ground truth for training. All we have during the training stage are the unlabeled videos with paired visual and audio components.

Cross-Modal Label Contrastive Learning with Expectation-Maximization

As illustrated in Fig 2, we propose a cross-modal label contrastive learning framework with expectation-maximization for unsupervised audio-visual event localization. The whole framework consists of three main steps: 1) At the feature representation level, the cross-modal collaborative learning aims to learn representations with two pretext tasks i.e. instance discrimination and feature decorrelation. 2) At the label level, a novel label contrasting pretext task trains the model to generate pseudo-labels with label contrasting between affinity and repulsion set. 3) An EM algorithm iteratively optimizes the pseudo-label acquisition and localization model to tackle the irrelevant background problem.

Feature-Level Cross-modal Collaborative Learning

Feature representation via cross-modal collaboration. The visual segments V_t and audio segments A_t are first processed by pretrained convolutional neural network. We denote visual and audio features processed by the pretrained network as $\{\boldsymbol{v}_t\}_{t=1}^{T}$ and $\{\boldsymbol{a}_t\}_{t=1}^{T}$, where $\boldsymbol{v}_t \in \mathbb{R}^{d_v \times (H*W)}$ and $\boldsymbol{a}_t \in \mathbb{R}^{d_a}$, and H, W are the height and width of the feature maps respectively. We follow the previous works (Xu et al. 2020) to design the network architecture of the audio-visual interaction. First, a multi-modal attention map $\boldsymbol{M}_t^s \in \mathbb{R}^{1 \times (H*W)}$ is generated by fusing visual and audio features with the attention mechanism as

$$\boldsymbol{M}_t^s = \text{softmax}\left(\boldsymbol{W}_{av}^s \left(\boldsymbol{W}_a^s \boldsymbol{a}_t \odot \boldsymbol{W}_v^s \boldsymbol{v}_t\right)\right) \tag{1}$$

where $\odot$ denotes Hadamard product, $\boldsymbol{W}_a^s \in \mathbb{R}^{d \times d_a}$ and $\boldsymbol{W}_v^s \in \mathbb{R}^{d \times d_v}$ are fully-connected layers with ReLU activation function for visual and audio features, respectively. $\boldsymbol{W}_{av}^s \in \mathbb{R}^{1 \times d}$ are learnable parameters with d as a hidden dimension. Then the spatial attentive features $\boldsymbol{v}_t^s \in \mathbb{R}^{d_v}$ are computed with the spatial attention map and the raw visual features $\boldsymbol{v}_t$ as

$$\boldsymbol{v}_t^s = \boldsymbol{M}_t^s \otimes \boldsymbol{v}_t^T, \tag{2}$$

where $\otimes$ denotes matrix multiplication.

To enable audio-visual interaction and collaborative learning from multi-modalities, we use two cross-attentional module, i.e. an audio-guided visual attention and a visual-guided audio attention module. The features of the two modalities $\boldsymbol{v}^s = \{\boldsymbol{v}_t^s\}_{t=1}^{T}$ and $\boldsymbol{a} = \{\boldsymbol{a}_t\}_{t=1}^{T}$ are first concatenated to $\boldsymbol{m}_{a,v} \in \mathbb{R}^{T \times 2d}$. The audio-collaborated visual features $\boldsymbol{v}_{attn} \in \mathbb{R}^{T \times d}$ are computed via multi-modal attention mechanism as

$$\boldsymbol{v}_{attn} = \text{Softmax}\left(\frac{\boldsymbol{Q}_v \boldsymbol{K}_v^T}{\sqrt{d}}\right) \boldsymbol{V}_v, \tag{3}$$

With learnable parameters $\boldsymbol{W}^Q \in \mathbb{R}^{2d\times T}, \boldsymbol{W}^K \in \mathbb{R}^{2d\times T}, \boldsymbol{W}^V \in \mathbb{R}^{d\times T}$, the above key features, value features, and query features are computed as

$$\boldsymbol{Q}_v = \boldsymbol{m}_{a,v}\boldsymbol{W}^Q, \boldsymbol{K}_v = \boldsymbol{m}_{a,v}\boldsymbol{W}^K, \boldsymbol{V}_v = \boldsymbol{v}^s\boldsymbol{W}^V. \quad (4)$$

Similarly, we can compute $\boldsymbol{Q}_a, \boldsymbol{K}_a, \boldsymbol{V}_a$, and the visual-collaborated audio features $\boldsymbol{a}_{attn} \in \mathbb{R}^{T\times d}$ are derived as

$$\boldsymbol{a}_{attn} = \text{Softmax}\left(\frac{\boldsymbol{Q}_a\boldsymbol{K}_a^T}{\sqrt{d}}\right)\boldsymbol{V}_a, \quad (5)$$

Self-supervised representation learning. Then we exploit two self-supervised pretext tasks for audio-visual representation learning, i.e. instance discrimination and feature decorrelation. For cross-modal collaborated features $\boldsymbol{v}_{attn}$ and $\boldsymbol{a}_{attn}$, we first apply average pooling to them separately and obtain $\boldsymbol{x}^v, \boldsymbol{x}^a \in \mathbb{R}^d$. The instance discrimination are defines as

$$\begin{aligned}\mathcal{L}_{ins} = &-\log\frac{\exp\left(x_+^a \cdot m_+^a/\tau\right)}{\sum_{i=1}^N \exp\left(x_+^a \cdot m_i^a/\tau\right)} \\ &-\log\frac{\exp\left(x_+^v \cdot m_+^v/\tau\right)}{\sum_{i=1}^N \exp\left(x_+^v \cdot m_i^v/\tau\right)}\end{aligned} \quad (6)$$

where τ is the temperature hyperparameter which we set it to 1.0. And the x_+^a, m_+^a and x_+^v, m_+^v are the positive pairs for audio/visual modality.. $\{m_i\}_{i=1}^N$ is the memory bank maintained with momentum mechanism for each modality.

The second self-supervised target is feature decorrelation, which makes each dimension of the learned representation to be orthogonal.

$$\mathcal{L}_{dec} = \left\|X^v(X^v)^T - I\right\|^2 + \left\|X^a(X^a)^T - I\right\|^2 \quad (7)$$

where x^v and x^a consists of each row of X^v and X^a. The final self-supervised loss objective function can be formulated as

$$\mathcal{L}_{self} = \mathcal{L}_{dec} + \mathcal{L}_{ins} \quad (8)$$

Self-Labeling via Label Contrastive Learning

To obtain pseudo-labels to train the audio-visual event localization model, we design a self-label annotator to automatically annotate the training dataset with pseudo-labels.

Construct affinity and repulsion set. After obtaining cross-modal representation f_i of each training sample as in section , its affinity set, and repulsion set are first constructed according to the similarity of the representations. We calculate the cosine similarity $s_{i,j}$ for each pair of training sample representation $\{f_i, f_j\}(1 \le i, j \le N)$. Then the top K most similar samples are assigned as its affinity set $\mathcal{A}_i$, and also set the top K most dissimilar samples as the repulsion set $\mathcal{R}_i$. And $\mathcal{A}_i, \mathcal{R}_i$ make up the contrastive set for label contrasting.

Self-label annotator with label contrasting. The self-label annotator consists of two layers of fully-connected layers. And it takes the cross-modal representation f_i as the input and ideally outputs one-hot labels which denote the event categories for the training samples. This pretext task of label contrasting. defines as narrowing the gap of pseudo-label distributions among the affinity set and vice versa distance it in the repulsion set. For each training sample f_i, a label affinitive term in the label contrasting loss encourages to reduce gap of the predicted pseudo-label $\Phi(f_i)$ to the pseudo-labels $\Phi(f_j)$ among its affinity set i.e. $j \in \mathcal{A}_i$. And on the contrary, the other label repulsion term to distance pseudo-label of training sample f_i from the pseudo-labels $\Phi(f_j)$ for the corresponding repulsion set $\mathcal{R}_i$. Specifically, the label contrastive loss is written as

$$\mathcal{L}_c = \sum_{j\in\mathcal{A}_i} \text{KL}\big(\Phi(f_i), \Phi(f_j)\big) - \sum_{j\in\mathcal{R}_i} \text{KL}\big(\Phi(f_i), \Phi(f_j)\big) \quad (9)$$

where KL denotes the Kullback–Leibler divergence of two distributions.

We further note that there exist trivial solutions with only minimizing the label contrasting term where parts of the pseudo-labels are empty and never assigned to the training samples. To tackle this issue, here we propose to regularize the pseudo-labels with entropy function $\mathcal{L}_e$:

$$\mathcal{L}_e = -\sum_{c=1}^{C} \overline{\Phi}(f)_c \log \overline{\Phi}(f)_c \quad (10)$$

where $\overline{\Phi}(f) = \frac{1}{N}\sum_{i=1}^N \Phi(f_i)$. With the regularization term, the final label contrastive loss $\mathcal{L}_{label}$ defines as the followings:

$$\mathcal{L}_{label} = \mathcal{L}_c + \mathcal{L}_e \quad (11)$$

in which the self-labeling generator can be trained in an end-to-end manner.

Expectation-Maximization Localization

Recall that expectation-maximization algorithm, is an elegant and powerful method for finding maximum likelihood solutions for model parameters with latent variables. In the E step, current values for the parameters are used to estimate the posterior probabilities of latent variables. Then these probabilities are then used in the M step to re-estimate the model parameters. In the scenarios of audio-visual event localization, on one hand, training an accurate localization model relies on pseudo-labels of high quality obtained from the audio-visual consensus event discovery. On the other hand, the visual and audio segments consist of both foreground events and background. These backgrounds, which are irrelevant to foreground audio-visual events as illustrated in Fig 1 (a), introduce noise to the event discovery process and thus hinder the acquisition of precise event categories by the pseudo-label generator. To tackle these issues, we regard the pseudo-label as the latent variables and then propose an expectation-maximization framework that embeds the label contrastive learning and audio-visual events localization.

Expectation step with self-label annotator. In the expectation step, we evaluate the latent pseudo-label distribution from the current values of localization model parameters. The audio and visual feature $a_{attn}, v_{attn} \in \mathbb{R}^{T\times d}$ from the collaborative representation learning are concatenated as

$\tilde{f}_i$. We first apply the localization model to get audio-visual events scores γ_{it}, where t denotes the t-th frame numbers for video i. The cross-modal representation f_i for video i is then computed with the estimated γ_{it} as

$$f_i = \sum_{t=1}^{T} \gamma_{it} \tilde{f}_{it} \tag{12}$$

With f_i based on the localization model, then posterior distribution of latent pseudo-label Z_i for video i are then estimated with label-contrastive learning as section :

$$Z_i = \Phi(f_i) \in \mathbb{R}^{C \times 1} \tag{13}$$

Note that for the first iteration of E step, the audio-visual event score γ can be initialized as $\gamma_{it} = \frac{1}{T}$ for all T frames, i.e. with the average localization results.

Maximization step with localization model. In the maximization step, we re-estimate the parameters of the localization model with the latent pseudo-label from the E step. After a cross-modal interaction module similar to Eq. 4 and cascaded to the self-supervised backbone, we can obtain the audio-visual feature $\boldsymbol{f}_{av}$ with dimensions of $T \times d_{av}$ as in (Xu et al. 2020). Then we predict two sorts of event scores for localization, i.e. an event-relevance score to tell which video segments belong to background or foreground, and an event category score to classify which category of the audio-visual event is performed in the segment. We compute the event-relevance score $\boldsymbol{\gamma} \in \mathbb{R}^{T \times 1}$ by applying a fully-connected layer $\boldsymbol{W}_\gamma$ cascaded with a softmax layers on $\boldsymbol{f}_{av}$, formulated as

$$\boldsymbol{\gamma} = \mathrm{Softmax}(\boldsymbol{f}_{av} \boldsymbol{W}_\gamma). \tag{14}$$

Similarly, we predict the event category score $\boldsymbol{s}_c \in \mathbb{R}^{T \times C}$ to identify event class for each segment as follows:

$$\boldsymbol{s}_c = \mathrm{Softmax}(\boldsymbol{f}_{av} \boldsymbol{W}_{av}). \tag{15}$$

where $\boldsymbol{W}_{av}$ is a fully-connected layer.

Note that the pseudo labels $\boldsymbol{Z}^{\text{unsup}}$ are assigned to the whole video and the category for each segment is not available. To tackle this issue, we formulate it as a Multiple Instance Learning (MIL) (Maron and Lozano-Pérez 1998) problem and exploit a MIL pooling layer to transform segment-level predictions s to video-level predictions $\boldsymbol{y} \in \mathbb{R}^{C \times 1}$, written as:

$$\boldsymbol{y} = \boldsymbol{s}_c^T \boldsymbol{\gamma} \tag{16}$$

Then we apply the binary cross-entropy (BCE) loss to video-level predictions and the pseudo video-labels as

$$\mathcal{L}_{MIL} = \mathcal{L}_{\mathrm{BCE}} \left(\boldsymbol{y}, \boldsymbol{Z}^{\text{unsup}} \right). \tag{17}$$

The localization model can then be trained in an end-to-end manner with Eq. 17. And the above E step and M step iteratively run until the convergence of the pseudo-label estimation and localization results.

Setting	Method	Feature	Acc
FS	AVEL (Tian et al. 2018)	VGG-19	68.6
	AVEL (Tian et al. 2018)	ResNet-151	74.7
	DAM (Wu et al. 2019)	VGG-19	74.5
	AVRB (Ramaswamy and Das 2020)	VGG-19	74.8
	AVIN (Ramaswamy 2020)	VGG-19	75.2
	AVSDN (Lin, Li, and Wang 2019)	ResNet-151	75.4
	CMRAN (Xu et al. 2020)	VGG-19	77.4
	CMRAN (Xu et al. 2020)	ResNet-151	78.3
	PSP (Zhou et al. 2021)	VGG-19	77.8
	M2N (Wang et al. 2021)	VGG-19	79.5
WS	AVEL (Tian et al. 2018)	VGG-19	66.7
	AVEL (Tian et al. 2018)	ResNet-151	73.3
	AVRB (Ramaswamy and Das 2020)	VGG-19	68.9
	AVIN (Ramaswamy 2020)	VGG-19	69.4
	AVSDN (Lin, Li, and Wang 2019)	ResNet-151	74.2
	AVT (Lin and Wang 2020)	VGG-19	70.2
	CMRAN (Xu et al. 2020)	VGG-19	72.9
	CMRAN (Xu et al. 2020)	ResNet-151	75.3
	PSP (Zhou et al. 2021)	VGG-19	73.5
US	Ours	VGG-19	63.2
	Ours	ResNet-151	67.1

Table 1: Performance comparison of localization results on the AVE dataset. FS, WS, and US represents the fully-supervised, weakly-supervised, and unsupervised settings.

Experiments

Dataset and Evaluation Metrics

AVE. Following existing works on fully/weakly-supervised settings (Zhou et al. 2021; Tian et al. 2018; Xu et al. 2020), we conduct our experiment on the AVE dataset (Tian et al. 2018). The AVE dataset is the standard datasets for audio-visual event localization, which contains 4143 video samples and 28 event categories that are collected from a wide domain of real-life scenes. We follow the identical setting to previous works for trainset/testset data splitting.

Evaluation metrics. We follow previous works (Xu et al. 2020) to use segment-level accuracy as the evaluation metrics for audio-visual event localization. To analyze the performance of the self-label annotator, we further adopt the metric of pseudo-label purity to evaluate the quality of the pseudo-labels, where we compute the largest percentage of samples with the same pseudo-label that also have the same ground-truth event category.

Implementation Details

Audio features and visual features. For audio representation, we apply the VGG-like network (Hershey et al. 2017) pretrained on AudioSet (Gemmeke et al. 2017) to extract acoustic features with dimensions of 128 for each audio segment. For fair comparison to existing fully/weakly-supervised works, we separately use the VGG-19 (Simonyan and Zisserman 2015) and ResNet-151 (He et al. 2016) pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2017) to extract the visual features. The visual features are with dimensions of $7 \times 7 \times 512$ and $7 \times 7 \times 2048$ for each segment respectively.

Loss function	Accuracy (%)
full loss	67.1
full w.o. L_{dec}	64.4
full w.o. L^{a}_{ins}	66.7
full w.o. L^{v}_{ins}	51.3

Table 2: Ablation study on collaborative learning loss.

Method	Accuracy (%)
full network	67.1
full w.o. s-attn	63.2
full w.o. audio2visual	59.0
full w.o. visual2audio	61.2

Table 3: Ablation study on collaborative learning modules.

Loss function	Label Purity	Accuracy (%)
full loss	77.81	67.1
full w.o. L_e	47.98	14.6
full w.o. L_c	12.3	7.6

Table 4: Ablation study on label contrastive loss.

EM step	Label purity	Accuracy (%)
1	76.90	65.8
2	77.35	66.7
3	77.81	67.1

Table 5: Localization results and label purity with EM step.

Modality	Accuracy (%)
audio	46.8
visual (resnet)	42.6
visual (vgg)	39.7
audio + visual (resnet)	67.1
audio + visual (vgg)	63.2

Table 6: Localization accuracy with various modalities.

The value of C	Accuracy (%)
28	67.1
56	66.7
84	66.1

Table 7: Localization accuracy with large C.

Training setup. The audio-visual collaborative learning and self-label contrastive model is trained with SGD optimizer with a learning rate of 0.05 and the batch size is set to 32. The learning rate is gradually decayed with a cosine decay schedule (Loshchilov and Hutter 2017). The hidden dimension d in the cross-modal collaboration module is set to 512. The number of parallel attention heads is set to 4. The memory bank is maintained with a momentum of 0.9 for each modality respectively. The numbers of training epochs for audio-visual collaborative learning and self-label contrastive model are both set to 200. To make the convergence more stable, we implement the feature decorrelation loss with its soft version as in (Tao, Takagi, and Nakata 2021). The hyperparameter C is set to 28. The contrastive set size K is set to 30. The localization model is trained with Adam (Kingma and Ba 2014) optimizer with a learning rate of 5×10^{-4} and weight decay of 5×10^{-4}. And the total EM step is set to 3.
Evaluation setup. Note that under the unsupervised setting, the event categories predicted by the model are based on C pseudo-label categories. To compare with previous methods, we assign each of the C pseudo-label categories with an event category in the AVE dataset. That is each pseudo-category is mapped to an event category with the largest frequency in this pseudo-category. Note that the pseudo-category assignment is only done for performance evaluation, and labels are strictly not needed during the training.

Performance Comparisons

Table 1 summarizes the localization accuracy under the fully-supervised, weakly-supervised, and unsupervised settings on the AVE dataset. Even without any video-level nor segment-level annotations, our unsupervised method can still achieve the competitive accuracy. For instance, the unsupervised method can achieve the competitive accuracy of 67.1% with ResNet-151 visual feature. This is about 85.7% and 89.1% of the accuracy of the best fully/weakly supervised counterparts. Also, with VGG-19 feature, the unsupervised localization result is with an accuracy of 63.2%, which is also comparative to previous settings requiring manual labeling (79.5% and 86.0% of the accuracy for the fully/weakly-supervised setting).

Ablation Study

1) Impact of cross-modal collaborative learning loss. We verify the impact of the cross-modal collaborative learning loss by ablation study on the AVE dataset. Table 2 summarizes localization accuracy when dropping different type loss terms. "full w.o. L_{dec}" refers to the model without using feature decorrelation loss, while "full w.o. L^{a}_{ins}" and "full w.o. L^{v}_{ins}" refers to the model dropping audio and visual parts of the instance discrimination loss. The accuracy drops about 2.7% without L_{dec}, which shows the benefit to decorrelate the audio-visual features. We also note that without either audio or visual part of the loss, the model accuracy would decrease with a clear margin.
2) Benefit of cross-modal collaborative learning modules. Table 3 summarizes ablation results of cross-modal collaborative learning modules, which show an evident performance drop without adopting the designed collaborative module. The method "full w.o. s-attn" replaces the audio-visual spatial attention by a mean pooling on the visual feature. And "full w.o. audio2visual" refers to the method which substitutes the audio-collaborated visual feature module with self-attention on single-modal visual feature, while "full w.o. visual2audio" is the method replacing visual-collaborated audio feature module with self-attention on audio feature.
3) Impact of label contrastive set size. Here we study impact of label contrastive set size on the localization accuracy. Fig. 4a presents the accuracy with respect to the varying con-

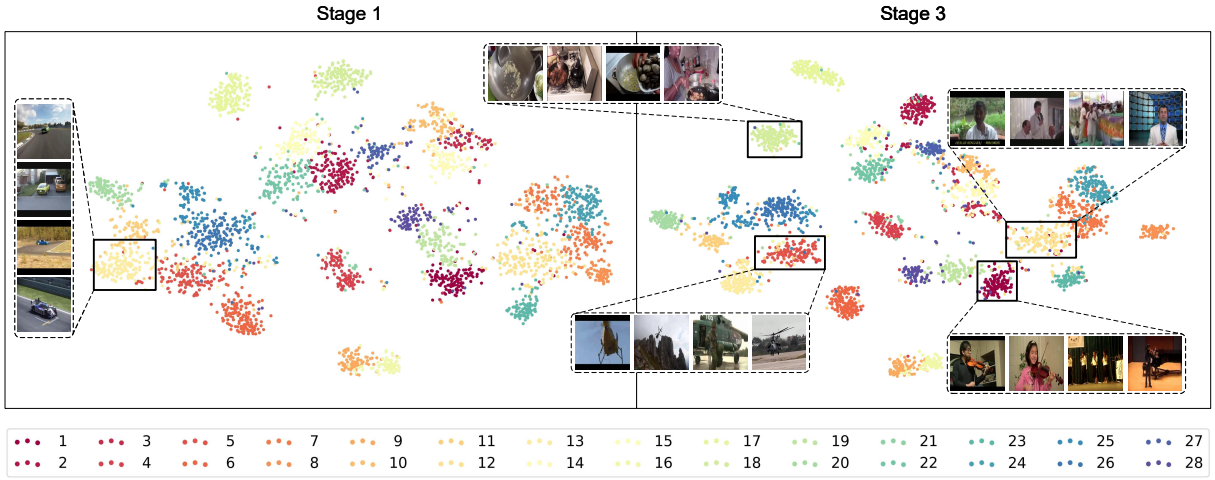
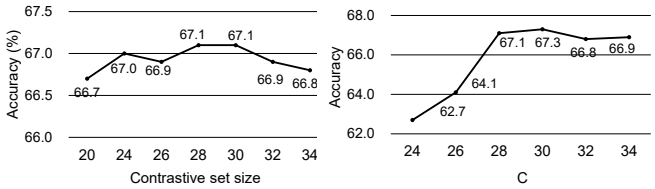


Figure 3: t-SNE visualization of pseudo-label estimation at the first and third EM-step. Different colors represent different event categories obtained by self-label annotator. Representative samples for the selected categories are shown in the dash boxes.



(a) Accuracy with contrastive set. (b) Accuracy with varying C.

Figure 4: Accuracy with contrastive set size and C.

trastive set size. We note that when contrastive set size is set to 28 or 30 achieves the best performance, although the model's accuracies are also satisfactory around them.

4) Impact of label contrastive loss. Table 4 illustrates ablation study results on the label contrastive loss. We drop the entropy term and label contrastive term from full loss function, and denote them as "full w.o. L_e" and "full w.o. L_c". Without either L_e or L_c, both the pseudo-label purity and localization accuracy degrade evidently. We further note that without L_e, the self-label annotator degenerates with only finding 12 pseudo categories and the other 16 are empty.

5) Effectiveness of expectation-maximization We summarize the pseudo-label purity and localization accuracy in Table 5 during the different EM steps. The label purity and localization accuracy consistently increase as the EM algorithm iterates. Figure 3 visualizes the pseudo-label distribution at the first and third EM steps with t-SNE algorithm. The visualization shows that the multi-modal features in the third step are better separated than in the first step.

6) The benefit of multi-modalities. Table 6 summarizes the localization accuracy with various modalities. We refer to the method with single modal input as "visual" and "audio' while denoting the method of modeling both modalities as "visual + audio". The "audio + visual" methods beat all single modal methods.

7) Impact of hyperparameter C. Fig. 4b summarizes the impact of the hyperparameter C. When C is smaller than the actual value 28, the accuracy drops evidently when C drops. However, when C becomes larger, the model performance keeps satisfactory (consistently above 66.5%), thanks to redundant event categories provided by a larger C. We further illustrate localization accuracy in Table 7 with C that is much larger than the ground-truth value, which indicates that a rough estimate of C over a wide range (larger than the ground-truth value) is sufficient.

Conclusions

In this paper, we tackle the unsupervised audio-visual event localization for the first time in the literature. We propose a cross-modal label contrastive learning framework that exploits cross-modal information among audio and visual modalities as self-supervision signals. An expectation-maximization algorithm is further devised to progressively optimize the pseudo-label acquisition and localization model. Extensive experiments show that our unsupervised approach achieves comparative performance to the state-of-the-art supervised counterparts.

Acknowledgements

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University (NTU), Singapore. The research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the NTU and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation) and A*STAR under it's A*STAR-P&G Joint Grant Call - DigiSolutions Accelerator Grant – Wave 3 (Award APG2013/138). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the A*STAR. This research work is also partially supported by the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

References

Alwassel, H.; Mahajan, D.; Korbar, B.; Torresani, L.; Ghanem, B.; and Tran, D. 2020. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. In *NeurIPS*.

Bulkin, D. A.; and Groh, J. M. 2006. Seeing sounds: visual and auditory interactions in the brain. *Current opinion in neurobiology*.

Cheng, Y.; Wang, R.; Pan, Z.; Feng, R.; and Zhang, Y. 2020. Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning. In *ACM MM*.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*.

Gong, G.; Wang, X.; Mu, Y.; and Tian, Q. 2020. Learning temporal co-attention models for unsupervised video action localization. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*.

Hu, D.; Nie, F.; and Li, X. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*.

Lin, Y.-B.; Li, Y.-J.; and Wang, Y.-C. F. 2019. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*.

Lin, Y.-B.; and Wang, Y.-C. F. 2020. Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization. In *ACCV*.

Loshchilov, I.; and Hutter, F. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*.

Maron, O.; and Lozano-Pérez, T. 1998. A framework for multiple-instance learning. *NeurIPS*.

Ramaswamy, J. 2020. What Makes the Sound?: A Dual-Modality Interacting Network for Audio-Visual Event Localization. In *ICASSP*.

Ramaswamy, J.; and Das, S. 2020. See the Sound, Hear the Pixels. In *WACV*.

Rouditchenko, A.; Zhao, H.; Gan, C.; McDermott, J.; and Torralba, A. 2019. Self-supervised audio-visual co-segmentation. In *ICASSP*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Smith, L.; and Gasser, M. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*.

Soomro, K.; and Shah, M. 2017. Unsupervised action discovery and localization in videos. In *CVPR*.

Tao, Y.; Takagi, K.; and Nakata, K. 2021. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *CILR*.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*.

Wang, H.; Zha, Z.-J.; Li, L.; Chen, X.; and Luo, J. 2021. Multi-Modulation Network for Audio-Visual Event Localization. *arXiv preprint arXiv:2108.11773*.

Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019. Dual attention matching for audio-visual event localization. In *ICCV*.

Xu, H.; Zeng, R.; Wu, Q.; Tan, M.; and Gan, C. 2020. Cross-Modal Relation-Aware Networks for Audio-Visual Event Localization. In *ACM MM*.

Xuan, H.; Zhang, Z.; Chen, S.; Yang, J.; and Yan, Y. 2020. Cross-Modal Attention Network for Temporal Inconsistent Audio-Visual Event Localization. In *AAAI*.

Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *CVPR*.