

國立臺灣大學電機資訊學院電信工程學研究所



碩士論文

Graduate Institute of Communication Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

利用模態獨立模型於弱監督式音頻-影像事件分析
Modality-Independent Teachers Meet Weakly-Supervised
Audio-Visual Event Parser

賴永玄

Yung-Hsuan Lai

指導教授：王鈺強 博士

Advisor: Yu-Chiang Frank Wang, Ph.D.

中華民國 112 年 7 月

July 2023



國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

利用模態獨立模型於弱監督式音頻-影像事件分析

Modality-Independent Teachers Meet Weakly-Supervised
Audio-Visual Event Parser

本論文係 賴永玄 (r10942097) 在國立臺灣大學電信工程學研究所完成之碩士學位論文，於民國 112 年 7 月 10 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Communication Engineering at National Taiwan University on July 10, 2023, has examined a Master's thesis entitled above presented by Yung-Hsuan Lai (r10942097) candidate and hereby certifies that it is worthy of acceptance.

口試委員 Oral examination committee:

王銀虎

(指導教授 Advisor)

孫品芳

陳祝嵩

系主任/所長 Director:

周錫堃

中文摘要



音頻-影像之跨模態學習是多模態機器學習中一個重要的研究領域。目前該領域主要聚焦於「模態對齊」的設定，即假設音頻和影像模態能「同時」提供預測目標的信號。然而在實際應用中，我們更常遇到「非模態對齊」的情況，即只能從音頻或影像「單一」模態中得到目標信號。為了深入研究非模態對齊情境，我們研究「音頻-影像影片事件分析」任務，此任務的目的是在非模態對齊且弱監督式學習的情境下，需識別影片中所有的聲音和影像事件（非模態對齊），並預測事件發生的時間。但是在訓練模型時，僅能使用影片層級的弱標籤（弱監督式學習），意即只能從此弱標籤知道影片中發生了哪些事件，但無法得知這些事件是經由哪種模態（聲音、影像或兩者同時）感知的，也無法確定事件發生的時間。為了應對這一挑戰，本研究提出一種簡單、有效且通用的方法——「影像-音頻標籤細化（VALOR）」。我們分別在音頻和影像模態引入大規模對比式預訓練模型作為模態教師模型，從而獲取含有事件模態和時間資訊的偽標籤，進一步對模型做偽標籤訓練。實驗結果顯示，VALOR 方法相較於基準方法使模型的平均 F-score 提升了 8.0。有趣的是，我們發現使用模態獨立的教師模型產生偽標籤在表現上優於使用模態融合的教師模型，這是因為前者能夠更好地抵抗非模態對齊的干擾。此外，我們的最佳模型在所有指標上都顯著優於前沿水平，使平均 F-score 提升了 5.4。最後，我們將 VALOR 方法推廣至「音頻-影像事件定位」任務，同樣在該任務上取得了超越其他方法和模型的最新成果，展現出卓越的普適性。





Modality-Independent Teachers Meet

Weakly-Supervised Audio-Visual

Event Parser

Yung-Hsuan Lai

Advisor: Yu-Chiang Frank Wang, Ph.D.

Graduate Institute of Communication Engineering

National Taiwan University

Taipei, Taiwan

July 2023





Abstract

Audio-visual learning has been a major pillar of multi-modal machine learning, where the community mostly focused on its *modality-aligned* setting, *i.e.*, the audio and visual modality are *both* assumed to signal the prediction target. With the Look, Listen, and Parse dataset (LLP), we investigate the under-explored *unaligned* setting, where the goal is to recognize audio and visual events in a video with only weak labels observed. Such weak video-level labels only tell what events happen without knowing the modality they are perceived (audio, visual, or both). To enhance learning in this challenging setting, we incorporate large-scale contrastively pre-trained models as the modality teachers. A simple, effective, and generic method, termed **V**isual-**A**udio **L**abel **E**laboration (VALOR), is innovated to harvest modality labels for the training events. Empirical studies show that the harvested labels significantly improve an attentional baseline by **8.0** in average F-score (Type@AV). Surprisingly, we found that modality-independent teachers outperform their modality-fused counterparts since they are noise-proof from the other potentially unaligned modality. Moreover, our best model achieves the new state-of-the-art on all metrics of LLP by a substantial margin (**+5.4** F-score for Type@AV). VALOR is further generalized to Audio-Visual Event Localization and achieves the new state-of-the-art as well.





Contents

Abstract	i
List of Figures	iv
List of Tables	vi
1 Introduction	1
2 Preliminary	5
2.1 Audio-Visual Video Parsing (AVVP)	5
2.2 Baseline Model	5
3 Proposed Method	7
3.1 Zero-Shot Transfer of Contrastive Pre-trained Models	7
3.2 Harvesting Training Signals	9
4 Related Work	11
4.1 Audio-Visual Video Parsing with Look, Listen, and Parse	11
4.2 More Audio-Visual Learning	12
5 Experiments	14
5.1 Experimental Setup	14
5.2 Unified Label Elaboration for State-of-the-Art Audio-Visual Video Parsing	16
5.3 Ablation Studies	17

CONTENTS

5.4 Qualitative Comparison with Previous AVVP Works	22
5.5 Generalize VALOR to Audio-Visual Event Localization	23
6 Conclusion	27
Reference	28





List of Figures

1.1	Modality-unaligned samples from LLP. Note that recent AVVP approaches like HAN [1] are vulnerable to unaligned data modality and produce incorrect predictions.	2
3.1	VALOR framework. With modality-independent label elaboration via CLIP and CLAP, the harvested temporally dense labels serve as additional modality- and time-aware cues for guiding HAN for AAVP.	8
5.1	Class-wise improvement on audio events. Comparing to the baseline model trained using video-level labels as audio segment-level labels, ours is trained on the derived audio segment-level pseudo labels.	21
5.2	Classwise improvement on visual events. Comparing to the baseline model trained using video-level labels as visual segment-level labels, ours is trained on the derived visual segment-level pseudo labels. Note that CLIP is applicable to extract global not fine-grained information from visual inputs. Thus, it is not expected to produce proper visual cues for the “Speech” event, which requires close attention to mouth movements.	21

LIST OF FIGURES

5.3 Qualitative comparison with previous AVVP works. In general, the predictions generated by our method VALOR are more accurate than those produced by the other methods. “GT” denotes the ground truth annotations. We compare with HAN [1] and JoMoLD [2].	22
--	----





List of Tables

5.1	Two different HAN model architectures. The “standard” model architecture is used in VALOR. The “variant” model architecture is used in VALOR+ and VALOR++.	15
5.2	AVVP benchmark. Note that pseudo label denoising is not applied for VPLAN [†] . VALOR+ is trained on a thinner yet deeper HAN of similar size. VALOR++ further uses CLIP and CLAP as feature extractors and significantly boosts all metrics.	16
5.3	Selection of modality-independent labeler. Note that utilizing a cross-modal labeler HAN instead of CLIP and CLAP to generate segment-level labels hardly improves the baseline (HAN). On the other hand, modality-less segment-level labels deteriorates the performance. All results are reported on the validation split of LLP.	18
5.4	The fidelity of our segment-level labels. We compare the segment-level labels generated from our method with the naive segment-level labels directly copied from the video-level labels, where we assume the events occurring in a video will occur in both modalities and every segment. We can see that the segment-level labels generated by our method VALOR are more accurate than the naive segment-level labels.	19



5.5 **Ablation study.** “global” denotes only video-level labels observed, while “dense” indicates segment-level labels available as ground truth. “base” is the baseline method [1]. “New Feat.” denotes the use of features from CLAP, CLIP, and R(2+1)D, and “Deep HAN” is that of the 256-dim 4-layer HAN model. All results are reported on the **validation** split of LLP. 20

5.6 **Results on the AVE task.** 26



Chapter 1

Introduction

Multi-modal learning has become a pivotal topic in modern machine learning research. Audio-visual learning is undoubtedly one of the primary focuses, as human frequently uses both hearing and vision to perceive the surrounding environment. Countless researchers have devoted to its *modality-aligned* setting with a strong assumption that the audio and visual modality *both* contain learnable clues to the desired prediction target. Numerous audio-visual tasks and algorithms have then been proposed, such as audio-visual speech recognition [3, 4, 5], audio-visual action recognition [6, 7, 8], sound generation from visual data [9, 10, 11], audio-visual question answering [12, 13], and many more. However, almost all real-world events can be audible while invisible, and *vice versa*, depending on how they are perceived. For example, a mother doing dishes in the kitchen might hear a baby crying from the living room, but be unable to directly see what is happening to the baby.

Having observed this potential modality mismatch in generic videos, Tian *et al.* [1] proposed the Audio-Visual Video Parsing (AVVP) task, which aims to recognize events in videos independently of the audio and visual modalities and also temporally localize these events. AVVP presents an *unaligned* setting of audio-visual learning since all 25 event types considered can be audio-only, visual-only, or audio-visual. Unfortunately, due to the laborious labeling process, Tian *et al.* [1]

1. Introduction

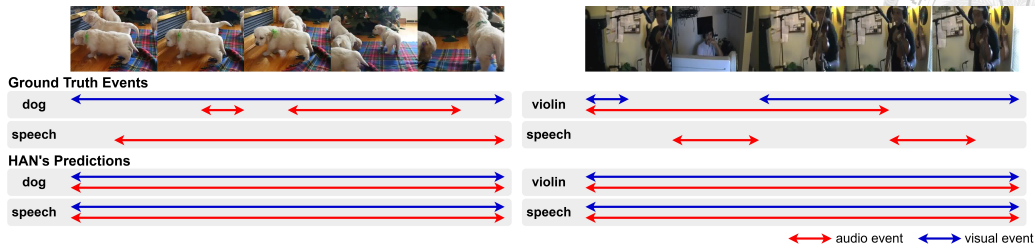


Figure 1.1: **Modality-unaligned samples from LLP.** Note that recent AVVP approaches like HAN [1] are vulnerable to unaligned data modality and produce incorrect predictions.

created this dataset (Look, Listen, and Parse; LLP) in a weakly-supervised setting.¹ More specifically, only video-level event annotations are available at training. In other words, the modality (audio, visual, or both) and the timestamp of which an event occurs are not given to the learning model.

The AVVP task poses significant challenges from three different perspectives. First, an event is typically modality independent, *i.e.*, knowing an event occurs in one modality says nothing about the other modality. As illustrated in Fig. 1.1, a sleeping dog is seen but may not be heard; conversely, a violin being played could sometimes go out of the camera view. Second, existing works heavily rely on the Multi-modal Multiple Instance Learning (MMIL) loss [1] to soft-select the modality (and timestamp), given only weak modality-less labels. This would be challenging for models to learn the correct event modality without observing a large amount of data. The uni-modal guided loss via label smoothing is also used to introduce uncertainty to the weak labels and thus regularize modality recognition. However, we hypothesize this improvement could be sub-optimal because no explicit modality information is introduced. Finally, AVVP requires models to predict events for all 1-second segments in a given video. Learning from weak video-level labels without timestamps makes it challenging for models to predict on a per-segment basis.

¹AVVP, LLP are used interchangeably in the literature. We use AVVP for the task, and LLP for the dataset.

To address the above challenges in AVVP, we propose to incorporate large-scale pre-trained open-vocabulary models, namely CLIP [14] and CLAP [15], to enhance learning with weak labels. Pre-trained on pixels and waveforms (and contrastively pre-trained with natural language), these models are inherently isolated from potential spurious noise from the other modality. Another benefit is the applicability of their prediction in an open-vocabulary fashion. Therefore, to benefit from CLIP and CLAP, we aim to harvest explicit modality learning signals from them. Moreover, we aim to inference these models per video segment, yielding fine-grained temporal annotations.

While it might be tempting to naively treat these pre-trained models as teachers and then applies knowledge distillation (KD) [16], this could be sub-optimal as some events are difficult to distinguish from a single modality, even for humans. For example, cars, motorcycles, and lawn mowers all produce similar sounds. To better utilize CLIP and CLAP, we introduce **Visual-Audio Label Elaboration (VALOR)**, to harvest modality and timestamp labels in LLP. We prompt CLIP/CLAP with natural language description of all visual/audio event types for each video segment-by-segment and then extract labels when a threshold is met. Additionally, implausible events are filtered out using the original weak labels accompanied with the video to mitigate the above indistinguishable problem. VALOR constructs fine-grained temporal labels in both modalities so that models have access to explicit training signals.

In addition to achieving the promising performance of AVVP, we observe that modality-independent teachers, CLIP and CLAP, generate more reliable labels than a modality-fused one, a cross-modal transformer. We also showcase the generalization capability of VALOR via the Audio-Visual Event Localization (AVE) task, in which our method also achieves the new state-of-the-art. Our contributions are summarized as follows:

- A simple and effective AVVP framework, VALOR, is proposed to harvest modality and temporal labels directly from video-level annotations, with an

1. Introduction

absolute improvement of **+8.0** F-score.

- We are the first to point out that modality independence could be crucial for audio-visual learning in the *unaligned* and weakly-supervised setup.
- Our VALOR achieves new state-of-the-art results with significant improvements on AVVP (**+5.4** F-score), with generalization to AVE (**+4.4** accuracy) jointly verified.





Chapter 2

Preliminary

2.1 Audio-Visual Video Parsing (AVVP)

The AVVP [1] task is to recognize events of interest in a video in both visual and audio modalities and to temporally identify the associated frames. For the benchmark dataset of Look, Listen, and Parse (LLP), a T -second video is split into T non-overlapping segments. Each video segment is paired with a set of multi-class event labels $(\mathbf{y}_t^v, \mathbf{y}_t^a) \in \{0, 1\}^C$ ($\mathbf{y}_t^v$: visual events, $\mathbf{y}_t^a$: audio events, C : number of event types). However, in the training split, the dense ‘segment-level’ labels $(\mathbf{y}_t^v, \mathbf{y}_t^a)$ are *not* available. Instead, only the global modality-less ‘video-level’ labels $\mathbf{y} := \max_t \{\mathbf{y}_t^v \wedge \mathbf{y}_t^a\}_{t=1}^T$ are provided ($\wedge$: element-wise ‘logical and’). In other words, AVVP models need to be learned in a weakly-supervised setting.

2.2 Baseline Model

We now briefly review the model of Hybrid Attention Network (HAN) [1], which is a common baseline for AVVP. In HAN, ResNet-152 [17] and R(2+1)D [18] are employed to extract 2D and 3D visual features. Subsequently, they are concatenated and projected into segment-level features $\mathbf{F}^v = \{\mathbf{f}_t^v\}_{t=1}^T \in \mathbb{R}^{T \times d}$ (d : hidden dimension). Segment-level audio features $\mathbf{F}^a = \{\mathbf{f}_t^a\}_{t=1}^T \in \mathbb{R}^{T \times d}$ are extracted



using VGGish [19] and projected to the same dimension. HAN takes these features and aggregates the intra-modal and cross-modal information through self-attention and cross-attention:

$$\tilde{\mathbf{f}}_t^a = \mathbf{f}_t^a + \text{Att}(\mathbf{f}_t^a, \mathbf{F}^a, \mathbf{F}^a) + \text{Att}(\mathbf{f}_t^a, \mathbf{F}^v, \mathbf{F}^v) \quad (2.1)$$

$$\tilde{\mathbf{f}}_t^v = \mathbf{f}_t^v + \text{Att}(\mathbf{f}_t^v, \mathbf{F}^v, \mathbf{F}^v) + \text{Att}(\mathbf{f}_t^v, \mathbf{F}^a, \mathbf{F}^a), \quad (2.2)$$

where $\text{Att}(\mathbf{q}, \mathbf{K}, \mathbf{V})$ denotes multi-head attention [20]. Following Transformer's practice, the outputs are further fed through LayerNorms [21] and a 2-layer FFN to yield $\hat{\mathbf{f}}_t^a, \hat{\mathbf{f}}_t^v$. With another linear layer, the hidden features are transformed into categorical logits $\mathbf{z}_t^v, \mathbf{z}_t^a$ for visual and audio events, respectively. Finally, the segment-level audio and visual event categorical probabilities, $\mathbf{p}_t^a$ and $\mathbf{p}_t^v$ ($\in [0, 1]^C$), are obtained by applying Sigmoid activation.

As a key module in Tian *et al.* [1], Multi-modal Multiple Instance Learning pooling (MMIL) is applied to address the above weakly-supervised learning task, which predicts the audio and visual event probabilities ($\mathbf{p}^m$, $m \in \{a, v\}$, audio and visual modalities) as:

$$\mathbf{A}^m = \{\boldsymbol{\alpha}_t^m\}_t = \text{softmax}_t(\hat{\mathbf{F}}^m \mathbf{W}^m), \quad \mathbf{p}^m = \sum_t \boldsymbol{\alpha}_t^m \odot \mathbf{p}_t^m, \quad (2.3)$$

where trainable parameters $\mathbf{W}^m \in \mathbb{R}^{d \times C}$ are implemented as linear layers ($\odot$: element-wise product). For video-level event probability $\mathbf{p}$:

$$\mathbf{B} = \{\{\boldsymbol{\beta}_t^m\}_t\}_m = \text{softmax}_m(\hat{\mathbf{F}} \mathbf{W}), \quad \mathbf{p} = \sum_m \sum_t \boldsymbol{\beta}_t^m \odot \boldsymbol{\alpha}_t^m \odot \mathbf{p}_t^m, \quad (2.4)$$

where $\hat{\mathbf{F}} = \{\hat{\mathbf{F}}^m\}_m \in \mathbb{R}^{2 \times T \times d}$ and $\mathbf{W}$ as a trainable linear layer. Moreover, modality training targets are obtained via label smoothing (LS) [22]: $\tilde{\mathbf{y}}^m = \text{LS}(\mathbf{y})$. Finally, the model is trained with binary cross entropy (BCE) as the loss function:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{guided}}^a + \mathcal{L}_{\text{guided}}^v, \quad \mathcal{L}_{\text{video}} = \text{BCE}(\mathbf{p}, \mathbf{y}), \quad \mathcal{L}_{\text{guided}}^m = \text{BCE}(\mathbf{p}^m, \tilde{\mathbf{y}}^m). \quad (2.5)$$

In summary, by the attention mechanisms introduced in HAN, MMIL pooling assigns event labels for each modality across time segments with only video-level event labels observed during training.



Chapter 3

Proposed Method

With only video-level event labels observed during training, we address three major challenges of AVVP: 1) modality independence of events' occurrence, 2) reliance on MMIL pooling for event label assignment under insufficient data, and 3) demand for dense temporal predictions. To address these challenges, we propose to leverage large-scale pre-trained contrastive models, CLIP and CLAP, to extract modality-aware, temporally dense training signals to guide model learning.

3.1 Zero-Shot Transfer of Contrastive Pre-trained Models

Radford *et al.* [14] proposed Contrastive Language-Image Pre-training (CLIP) to utilize web-scale image-text pairs to train a strong image encoder. As a result, CLIP overthrows the limitation of predicting predefined categories. Due to its large training data size (400M), CLIP has demonstrated remarkable zero-shot performance on a wide range of visual recognition tasks. All the above motivates us to incorporate CLIP to improve visual event recognition in AVVP.

In our work, CLIP's visual understanding of AVVP is extracted as the following. We extract T evenly spaced video frames and pass them into CLIP's image encoder to obtain the visual features $\{\mathbf{f}_t^{\text{CLIP}}\}_{t=1}^T \in \mathbb{R}^{T \times d_2}$ (d_2 : the dimension of CLIP's

3. Proposed Method

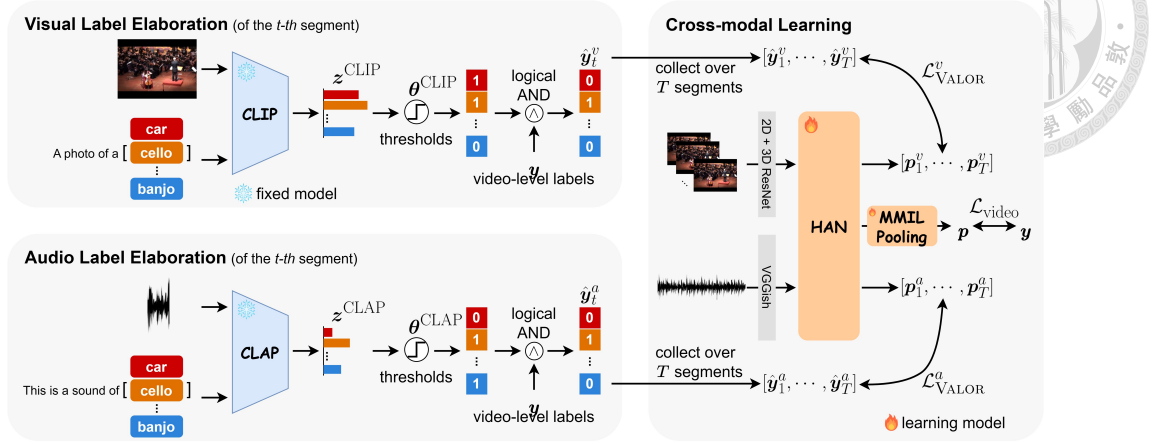


Figure 3.1: **VALOR framework.** With modality-independent label elaboration via CLIP and CLAP, the harvested temporally dense labels serve as additional modality- and time-aware cues for guiding HAN for AAVP.

feature). For simplicity and readability, we will omit the time subscript t for the remainder of this paper when there is no ambiguity. Next, we convert the AVVP event categories to concepts that CLIP understands. A caption for each event is constructed by adding a “A photo of” prefix to the event’s natural language form. These captions are processed by the CLIP’s text encoder, resulting in event features $\mathbf{G}^{\text{CLIP}} = \{\mathbf{g}_c^{\text{CLIP}}\}_{c=1}^C \in \mathbb{R}^{C \times d_2}$, where c indexes the events, and $\mathbf{g}_c$ represents the text feature of the c -th event. Frame-level event logits $\mathbf{z}^{\text{CLIP}} \in \mathbb{R}^C$ can be obtained by calculating the inner products:

$$\mathbf{z}^{\text{CLIP}} = \mathbf{f}^{\text{CLIP}} \mathbf{G}^{\text{CLIP}^\top}. \quad (3.1)$$

In light of the notable success of CLIP [14], several studies have sprouted to research on learning representative audio embeddings and text embeddings through Contrastive Language-Audio Pretraining (CLAP) [23, 24, 25, 26, 15]. In the same way as images and text are encoded in CLIP, web-scale audios and text are pre-trained with a contrastive objective in CLAP. Symmetrically, we obtain CLAP’s understanding of AVVP audios as the following. From the audio channel, the raw waveform is extracted and split into T segments with the same lengths and then fed into CLAP, yielding segment-level audio features $\{\mathbf{f}_t^{\text{CLAP}}\}_t \in \mathbb{R}^{T \times d_3}$ (d_3 :

the dimension of CLAP’s feature). On the other hand, an audio event caption is constructed by adding the prefix “This is a sound of” to each AVVP event’s name. Processed by the CLAP text encoder, we obtain $\mathbf{G}^{\text{CLAP}} = \{\mathbf{g}_c^{\text{CLAP}}\}_c \in \mathbb{R}^{C \times d_3}$. Segment-level audio event logits $\mathbf{z}^{\text{CLAP}} \in \mathbb{R}^C$ are obtained by the inner products:

$$\mathbf{z}^{\text{CLAP}} = \mathbf{f}^{\text{CLAP}} \mathbf{G}^{\text{CLAP}^\top}. \quad (3.2)$$

We note that Eqn. (3.1) and (3.2) can be viewed as CLIP’s and CLAP’s understanding of the associate video frame in the event space of AVVP.

3.2 Harvesting Training Signals

Given the logits $\mathbf{z}^{\text{CLIP}}$ and $\mathbf{z}^{\text{CLAP}}$, we aim to convert them to useful training signals for the AVVP task. An intuitive idea is to teach our model via knowledge distillation (KD) [16]. To deploy KD in training, segment-level normalized probabilities are first computed: $\mathbf{q}^P = \text{softmax}_c(\mathbf{z}^P)$, $\mathbf{q}^m = \text{softmax}_c(\mathbf{z}^m)$, where $(m, P) \in \{(v, \text{CLIP}), (a, \text{CLAP})\}$ denotes data modality (audio/visual) and pre-trained model (CLIP/CLAP) pair. Next, KL-divergence for all segments is calculated: $\mathcal{L}_{\text{KD}}^m = \sum_t \text{KL}(\mathbf{q}_t^P, \mathbf{q}_t^m)$. Finally, KD training is done by optimizing the loss function:

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{KD}}^a + \mathcal{L}_{\text{KD}}^v. \quad (3.3)$$

However, as we find out empirically (shown in Table 5.5), this is *not* the optimal usage of CLIP and CLAP. We hypothesize that some events are hard to distinguish from a single modality, *e.g.* cars, motorcycles, and lawn mowers produce the sound of an engine. Therefore, we design VALOR, utilizing video-level labels to filter out the impossible events, hence mitigating the confusion.

Visual-Audio Label Elaboration (VALOR). To better exploit CLIP and CLAP, we design a simple yet effective method, VALOR, to harvest dense labels in both modalities. In particular, we first define class-dependent thresholds $\boldsymbol{\theta}^P \in \mathbb{R}^C$. The segment-level labels for each modality are further obtained by thresholding the

3. Proposed Method

contrastive models understanding logits $\hat{\mathbf{y}}_t^m = \mathbf{y} \wedge \{\mathbf{z}_t^P > \boldsymbol{\theta}^P\}$. The overall loss function can be written as:

$$\mathcal{L}_{\text{VALOR}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{VALOR}}^a + \mathcal{L}_{\text{VALOR}}^v, \quad \mathcal{L}_{\text{VALOR}}^m = \sum_t \text{BCE}(\mathbf{p}_t^m, \hat{\mathbf{y}}_t^m). \quad (3.4)$$

To summarize, we design a simple yet effective method, VALOR, to utilize large-scale pre-trained contrastive models, CLIP and CLAP, to generate segment-level labels in both modalities. Due to the nature of immunity to spurious noise from the other modality, the contrastive pre-training methods, and the large pre-training dataset size, CLIP and CLAP are able to provide reliable labels in visual and audio modality, respectively. In addition, they are able to provide temporally dense labels to explicitly guide the model in learning events in each segment.





Chapter 4

Related Work

4.1 Audio-Visual Video Parsing with Look, Listen, and Parse

For AVVP, research flourishes along two orthogonal directions: enhancing the model architecture and label refinement. Architectural improvements include cross-modal co-occurrence module [27], class-aware uni-modal features and cross-modal grouping [28], and Multi-Modal Pyramid attention [29]. On the other hand, label refinement shares a similar spirit with ours. MA [30] corrupted the data by swapping the audio channel of two videos with disjoint video-level event sets. The model’s likelihood of the corrupted data was then used to determine the modality label. More recently, JoMoLD [2] utilized a two-stage approach. First, an AVVP model was trained as usual. Next, another model was trained while denoising the weak labels with prior belief from the first model. Both MA and JoMoLD produced global modality labels without timestamps. Concurrent to ours, VPLAN [31] generates dense temporal visual annotations with CLIP; however, the audio labels remain absent. Our VALOR represents a *unified* framework to elaborate the weak labels, along modality *and* temporal dimension, via zero-shot transfer of pre-trained models. We further emphasize the importance of modality



independence when synthesizing modality supervision.

4.2 More Audio-Visual Learning

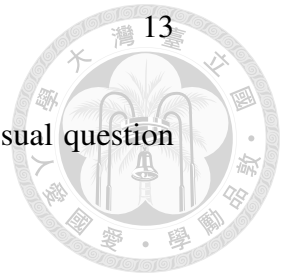
Audio-Visual Event Localization (AVE). Tian *et al.* [32] proposed AVE to recognize the audio-visual event in a video while localizing its temporal boundaries. Numerous studies have been conducted, including Lin *et al.* [33] with seq2seq models, Lin and Wang [34] using intra&inter frame Transformers, Wu *et al.* [35] via dual attention matching, audio-spatial channel-attention by Xu *et al.* [36], positive sample propagation from Zhou *et al.* [37], and Xia and Zhao [38] employing background suppression. We generalize VALOR to AVE's weakly supervised setting.

Audio-Visual Assistance. While significant advancements have been made in speech recognition, speech enhancement, and action recognition, noise or bias residing in the uni-modal data is still problematic. An effective solution could involve integrating data from an additional modality. This research direction encompasses various areas including speech recognition [39, 3, 4, 5], speaker recognition [40, 41, 42, 43, 44, 45], action recognition [46, 6, 47, 7, 8], speech enhancement or separation [48, 49, 50, 51, 52, 53], and object sound separation [54, 55, 56, 57, 58, 59, 60, 61].

Audio-Visual Correspondence and Understanding. Humans possess an impressive capacity to deduce occurrences in one sensory modality using information solely from another. This fascinating human ability to perceive across modalities has inspired researchers to delve into sound generation from visual data [9, 62, 10, 63, 64, 65, 11, 27], video generation from audio [66, 67, 68, 69, 70], and audio-visual retrieval [71, 72]. In the pursuit of understanding how humans process audio-visual events, numerous studies have been undertaken on audio-visual understanding tasks such as sound localization in videos [73, 74, 75, 76, 77],

4. *Related Work*

audio-visual navigation [78, 79, 80, 81, 82, 83, 84], and audio-visual question answering [85, 86, 87, 88, 12, 13, 89].





Chapter 5

Experiments

5.1 Experimental Setup

Dataset and Metrics. The LLP dataset is composed of 11849 10-second Youtube video clips covering 25 event categories, such as human activities, musical instruments, vehicles, and animals. The dataset is divided into training, validation, and testing splits, containing 10,000, 649, and 1200 clips, respectively. The official evaluation uses F-score to evaluate audio (A), visual (V), and audio-visual (AV) events separately. Type@AV (Type) is the averaged F-scores of A, V, and AV. Event@AV (Event) measures the ability to detect events in both modalities by combining audio and visual event detection results. Different from segment-level metrics, the event-level metrics treat consecutive positive segments as a whole, and mIoU of 0.5 is applied to calculate F-scores.

Implementation Details. Unless otherwise specified, VALOR uses HAN under a fair setting w.r.t. previous works with same data pre-processing. For the visual feature extraction, video frames are sampled at 8 frames per second. Additionally, we conduct experiments using CLIP and CLAP as feature extractors. The pre-trained ViT-L CLIP and HTSAT-RoBERTa-fusion CLAP are used to generate labels and extract features. Note that for all experiments with CLAP, we use the

Table 5.1: **Two different HAN model architectures.** The “standard” model architecture is used in VALOR. The “variant” model architecture is used in VALOR+ and VALOR++.

HAN model	standard	variant
Model Arch. Hyper-parameters		
hidden dim	512	256
hidden layers	1	4
trainable params	5.1M	5.05M
Training Hyper-parameters		
peak learning rate	1e-4	3e-4
min learning rate	1e-6	3e-6

implementation from Wu *et al.* [15] pre-trained on LAION-Audio-630K. We do not use the version pre-trained on AudioSet (a larger pre-training corpus) since it overlaps with the AVVP validation and testing videos.

In our experiments, we apply two different model architectures: 1) the standard model architecture, which is employed in VALOR, consists of a single HAN layer with a hidden dimension of 512; 2) the variant model architecture, which is used in VALOR+ and VALOR++, is a thinner yet deeper HAN model, comprising four HAN layers with a hidden dimension of 256. Both models contain approximately the same number of trainable parameters. The above details are summarized in Table 5.1. The models are trained using the AdamW optimizer, configured with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay set to 0.001. We employ a learning rate scheduling approach that initiates with a linear warm-up phase over 10 epochs, rises to the peak learning rate, and then decays according to a cosine annealing schedule to the minimum learning rate. We set the batch size to 64 and train for 60 epochs in total. We clip the gradient norm at 1.0 during training.

Table 5.2: **AVVP benchmark.** Note that pseudo label denoising is not applied for VPLAN[†]. VALOR+ is trained on a thinner yet deeper HAN of similar size. VALOR++ further uses CLIP and CLAP as feature extractors and significantly boosts all metrics.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVE [32]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5	36.5
AVSDN [33]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [1]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Pyr [29]	60.9	54.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.5
MGN [28]	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
CVCMS [90]	59.2	59.9	53.4	57.5	58.1	51.3	55.5	46.2	51.0	49.7
DHHN [91]	61.3	58.3	52.9	57.5	58.1	54.0	55.1	47.3	51.5	51.5
MA [30]	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
JoMoLD [2]	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
VPLAN [†] [31]	60.5	64.8	58.3	61.2	59.4	51.4	61.5	51.2	54.7	50.8
VALOR	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
VALOR+	<u>62.8</u>	<u>66.7</u>	<u>60.0</u>	<u>63.2</u>	<u>62.3</u>	<u>57.1</u>	<u>63.9</u>	<u>54.4</u>	<u>58.5</u>	<u>55.9</u>
VALOR++	68.1	68.4	61.9	66.2	66.8	61.2	64.7	55.5	60.4	59.0

5.2 Unified Label Elaboration for State-of-the-Art Audio-Visual Video Parsing

To demonstrate the effectiveness of VALOR, we evaluate our method on the AVVP benchmark. Existing works include: 1) weakly-supervised audio-visual event localization methods AVE and AVSDN, 2) HAN and its network architecture advancements MM-Pyramid, MGN, CVCMS, and DHHN, and 3) different label refinement methods MA, JoMoLD, and VPLAN. We report the results on the test split of the LLP dataset in Table 5.2.

We achieve the new state-of-the-art (SOTA) on all metrics consistently with a large margin. Our method VALOR significantly improves the baseline (HAN) by **8.0** in segment-level Type@AV. Compared to previous published SOTA, JoMoLD, VALOR scores higher on all metrics, including the **5.4** F-score improvement for segment-level Type@AV, under a fair setting. With light hyper-parameter tuning, VALOR+ further achieves a significant 2.4 improvement on Type@AV, with a deeper yet thinner HAN while keeping a similar number of trainable parameters. Our improvement on the audio side w.r.t. the concurrent preprint VPLAN is more significant than the visual side, which may be attributed to our effective audio teacher CLAP and label elaboration along the modality axis. We empirically conclude that VALOR has successfully unified label refinement along both modality and temporal dimensions. To push to the limits, we further proposed VALOR++ by replacing the feature extraction models with CLIP and CLAP, achieving another consistent boost, including 3.0 in segment Type@AV. We will release the VALOR++ pre-trained checkpoint, features, and harvested labels to boost future AVVP research.

5.3 Ablation Studies

The impressive results achieved in Table 5.2 are based on careful design. In this subsection, we elaborate on why we choose CLIP and CLAP to synthesize dense labels with modality annotations with empirical support. Furthermore, we break down the loss function and modeling components into orthogonal pieces and evaluate their individual effectiveness.

How to choose the labeler?. In Table 5.3, we show the necessity of modality-independent pre-trained models (CLIP and CLAP) over the multi-modal model (HAN) as the labeler (2nd row) and that modality-aware labels beat modality-agnostic labels (3rd row). We aim to demonstrate the necessity and importance of **using large-scale pre-trained uni-modal models** to annotate **modality-aware**

Table 5.3: **Selection of modality-independent labeler.** Note that utilizing a cross-modal labeler HAN instead of CLIP and CLAP to generate segment-level labels hardly improves the baseline (HAN). On the other hand, modality-less segment-level labels deteriorates the performance. All results are reported on the **validation** split of LLP.

Dense Labeler	Modality Label	Segment-level					Event-level				
		A	V	AV	Type	Event	A	V	AV	Type	Event
None	✓	62.0	54.5	50.2	55.6	57.1	53.5	50.5	43.6	49.2	50.3
HAN	✓	62.1	56.4	52.1	56.8	57.6	53.4	52.0	45.4	50.3	50.6
CLIP&CLAP	✗	41.0	59.0	34.5	44.9	52.1	33.2	56.2	28.2	39.2	43.1
CLIP&CLAP	✓	62.7	66.3	61.0	63.4	61.8	55.5	62.0	54.1	57.2	53.8

segment-level labels. To validate the former, we employ a baseline model (HAN) that has been trained on AVVP to individually annotate segment-level labels within the two modalities. Experimental results show that modality-aware temporal dense labels generated by a multi-modal model (HAN), learned from weak labels, are less effective than those generated by large-scale pre-trained uni-modal models (CLIP and CLAP), thereby underscoring the essentiality of using large-scale pre-trained uni-modal models. Subsequently, to validate the latter, we generate modality-agnostic segment-level labels from CLIP and CLAP, meaning that these labels only reveal the events occurring in each segment but do not disclose the modality of the event. As seen from the third row of Table 5.3, while such a labeling method increases the F-score for visual events, it dramatically decreases the F-score for audio events. The overall performance (Type F-score) is even worse than that of HAN (the first row), clearly indicating the importance of modality-aware labeling for the model to learn the AVVP task effectively.

How accurate are our segment-level labels?. To examine the fidelity of our generated segment-level pseudo labels in both modalities, we compare our labels, $\hat{y}_t^a$ and $\hat{y}_t^v$, with naive segment-level labels, which are obtained by replicating

Table 5.4: **The fidelity of our segment-level labels.** We compare the segment-level labels generated from our method with the naive segment-level labels directly copied from the video-level labels, where we assume the events occurring in a video will occur in both modalities and every segment. We can see that the segment-level labels generated by our method VALOR are more accurate than the naive segment-level labels.

Methods	Audio	Visual	Audio-Visual
video labels as segment labels	79.33	69.30	60.69
VALOR-generated segment labels	84.92 (+5.59)	82.80 (+13.50)	76.37 (+15.68)

the video-level label y of a video $2T$ times and assigning the duplicates to all segments in the video in both modalities. In other words, we assume that an event occurs in both modalities and all segments if it occurs in the video. As depicted in Table 5.4, the segment-level F-scores of our generated segment-level audio and visual pseudo labels are superior to those of the respective naive ones. Notably, our segment-level visual F-score is 10 points higher than the naive one. Moreover, we evaluate the fidelity of the audio-visual event labels by performing element-wise AND operation on the segment-level audio and visual labels. The segment-level F-score of our audio-visual labels significantly surpasses that of the naive ones. These findings present the reliability of our segment-level pseudo labels, which can provide more accurate segment-level information to facilitate model training.

How to use the elaborated labels? Finally, we conduct an ablation study on utilizing CLIP and CLAP together. The results are presented in Table 5.5. The replacement of the smoothed video-level event labels $\tilde{y}^a$ and $\tilde{y}^v$ with their respective refined weak labels $\hat{y}^a$ and $\hat{y}^v$ derived from our method lead to a significant enhancement in the Type@AV F-score, from 54.0 to 60.8. This finding underscores the importance of incorporating labels that are proximal to ground truth, albeit weak. Furthermore, we leverage the CLIP and CLAP models to generate segment-level

Table 5.5: **Ablation study.** “global” denotes only video-level labels observed, while “dense” indicates segment-level labels available as ground truth. “base” is the baseline method [1]. “New Feat.” denotes the use of features from CLAP, CLIP, and R(2+1)D, and “Deep HAN” is that of the 256-dim 4-layer HAN model. All results are reported on the **validation** split of LLP.

Audio Loss		Visual Loss		New	Deep	Segment-level				
global	dense	global	dense	Feat.	HAN	A	V	AV	Type	Event
base	✗	base	✗	✗	✗	62.0	54.5	50.2	55.6	57.1
✗	KD	✗	KD	✗	✗	51.1	64.0	48.0	54.3	55.5
VALOR	✗	VALOR	✗	✗	✗	62.1	65.8	59.0	62.3	61.2
base	✗	✗	VALOR	✗	✗	60.5	66.7	60.8	62.7	59.8
✗	VALOR	base	✗	✗	✗	62.2	54.5	52.7	56.5	56.5
✗	VALOR	✗	VALOR	✗	✗	62.7	66.3	61.0	63.4	61.8
✗	VALOR	✗	VALOR	✗	✓	64.5	67.1	63.1	64.9	63.2
✗	VALOR	✗	VALOR	✓	✓	71.4	69.4	64.9	68.6	69.7

labels for each modality. This approach results in an improvement of 8.0 Type@AV F-score over the baseline method, indicating that explicitly informing the model of the events occurring in each segment of the audio-visual video facilitates the learning of the Audio-Visual Video Parsing (AVVP) task. In addition to guiding the model, CLIP and CLAP were also used to obtain more representative features. Replacing the ResNet-152 and VGGish features with CLIP and CLAP features yielded a Type@AV F-score improvement of 4.0.

Classwise F-score Comparison.. We further evaluate the effectiveness of providing accurate uni-modal segment-level pseudo labels for the model training. We visualize classwise improvements between our generated segment-level labels for each modality and the naive segment-level labels derived from the video-level labels. In Figure 5.1, we observe that when our audio pseudo labels are used, most of the audio events improve. In Figure 5.2, when our visual pseudo labels are used,

5. Experiments

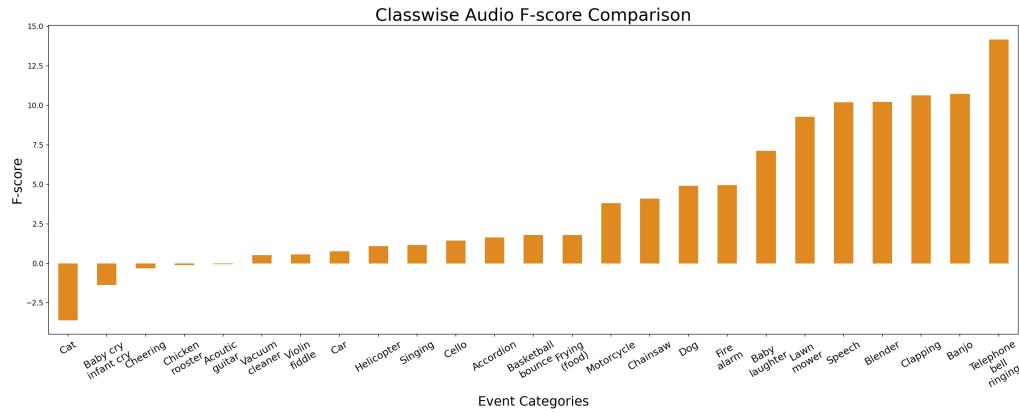


Figure 5.1: **Class-wise improvement on audio events.** Comparing to the baseline model trained using video-level labels as audio segment-level labels, ours is trained on the derived audio segment-level pseudo labels.

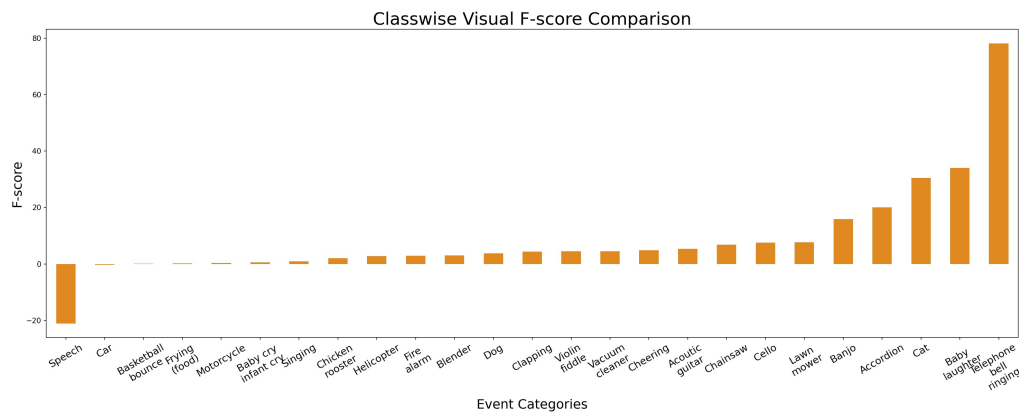


Figure 5.2: **Classwise improvement on visual events.** Comparing to the baseline model trained using video-level labels as visual segment-level labels, ours is trained on the derived visual segment-level pseudo labels. Note that CLIP is applicable to extract global not fine-grained information from visual inputs. Thus, it is not expected to produce proper visual cues for the “Speech” event, which requires close attention to mouth movements.

nearly every event’s F-score increases. These results indicate the effectiveness of our method in guiding the model to learn events in each modality. For the inferior performance on the “Speech” event, since CLIP is inept at extracting fine-grained visual information, it is not expected to recognize the “Speech” event well, which

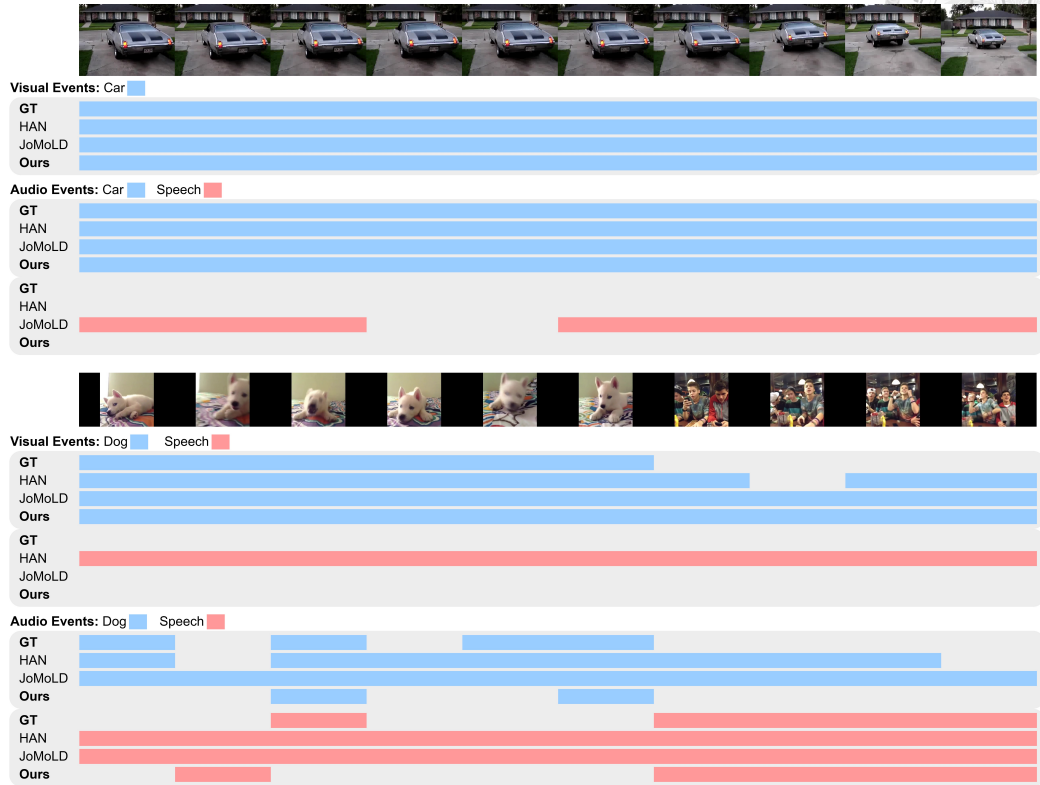


Figure 5.3: **Qualitative comparison with previous AVVP works.** In general, the predictions generated by our method VALOR are more accurate than those produced by the other methods. “GT” denotes the ground truth annotations. We compare with HAN [1] and JoMoLD [2].

requires close attention on mouth movements.

5.4 Qualitative Comparison with Previous AVVP Works

Aside from quantitative comparison with previous AVVP works, we perform a qualitative evaluation as well. In Figure 5.3, we qualitatively compare with the baseline method HAN [1] and the state-of-the-art method JoMoLD [2]. In the top video example, JoMoLD erroneously predicts the “Speech” audio event, while all other methods accurately identify the audio events. In the bottom example,

HAN produces identical temporal annotations for the “Speech” event in both modalities, despite the event only occurring audibly. Additionally, our method provides annotations that more closely align with the ground truth than either HAN or JoMoLD does when the events occur intermittently, which is challenging for models to generate accurate predictions.

5.5 Generalize VALOR to Audio-Visual Event Localization

In this section, we showcase the additional generalization ability of VALOR by applying it to the Audio-Visual Event Localization (AVE) task. We consider the weakly-supervised version of AVE, where segment-level ground truths are not available to the model during training, meaning that no timestamp is provided for the event, motivating us to apply VALOR to harvest event labels. Without task-specific modification, we directly apply HAN and VALOR to AVE. The only difference is that at inference, we combine the audio and visual prediction to obtain the audio-visual event required in this task.

Baseline Method. We adopt the baseline model HAN to aggregate uni-modal and cross-modal temporal information as we have done in the AVVP task. For brevity, we introduce our baseline method from the procedure after feature aggregation. The aggregated segment-level audio features and visual features, $\hat{f}_t^a$ and $\hat{f}_t^v$ ($\in \mathbb{R}^d$), output from HAN are processed through a 2-layer feed-forward network (FFN) to yield the uni-modal segment-level logits, z_t^a and z_t^v ($\in \mathbb{R}^{(C+1)}$), respectively:

$$z_t^m = \text{FFN}(\hat{f}_t^m), \quad m \in \{a, v\}, \quad (5.1)$$

where $C+1$ denotes the number of event classes and the “background” event. Since segment-level labels are not available in the weakly-supervised setting, we simply infer video-level logits $z \in \mathbb{R}^{C+1}$ by averaging all logits over time dimension t and

modality dimension m . Finally, the binary cross-entropy loss is applied to train the model:

$$\mathcal{L}_{\text{video}}^{\text{ave}} = \text{BCE}(\text{Sigmoid}(z), y), \quad z = \frac{1}{2T} \sum_t \sum_m z_t^m \quad (5.2)$$

Harvesting Training Signals. The main idea of our method VALOR is to leverage large-scale pre-trained contrastive models to provide reliable temporally dense labels in both modalities. We elaborate on how these labels are generated. Initially, segment-level audio logits and visual logits, z_t^{CLAP} and $z_t^{\text{CLIP}} (\in \mathbb{R}^C)$, are generated in a manner identical to the AVVP task. We then construct the uni-modal segment-level labels $\hat{y}_t^a$ and $\hat{y}_t^v (\in \mathbb{R}^C)$ by comparing the logits with the pre-defined class dependent thresholds ϕ^{CLAP} and $\phi^{\text{CLIP}} (\in \mathbb{R}^C)$:

$$\hat{y}_t^m = y \wedge \{z_t^P > \phi^P\}, \quad (m, P) \in \{(v, \text{CLIP}), (a, \text{CLAP})\} \quad (5.3)$$

In addition, we append an additional event “background” to the end of the segment-level labels $\hat{y}_t^m$ to expand the dimension to $\mathbb{R}^{C+1}$. If the original $\hat{y}_t^m$ consists solely of zeros, we assign the last dimension (the “background” event) a value of one; otherwise, we assign it a zero. Having prepared the temporally dense labels in both modalities $\hat{y}_t^a$ and $\hat{y}_t^v$, we apply binary cross-entropy loss in individual modality and combine them to optimize the whole model instead of using the video-level loss $\mathcal{L}_{\text{video}}^{\text{ave}}$:

$$\mathcal{L}_{\text{VALOR}}^{\text{ave}} = \text{BCE}(\text{Sigmoid}(z_t^a), \hat{y}_t^a) + \text{BCE}(\text{Sigmoid}(z_t^v), \hat{y}_t^v) \quad (5.4)$$

Dataset & Evaluation Metrics. The *Audio-Visual Event (AVE) Dataset* [32] is composed of 4143 10-second video clips from AudioSet [92] that cover 28 real-world event categories, such as human activities, musical instruments, vehicles, and animals. Each clip contains an event and is uniformly split into ten segments. Each segment is annotated with an event category if the event can be detected through both visual and auditory cues; otherwise, the segment is labeled as background. However, we only use video-level labels indicating which event occurs in the video

during training. We follow [32] to split the AVE dataset into training, validation, and testing split and report the results on the testing split. Following the previous work [32], we use the accuracy of segment-level event category predictions as the evaluation metric.

Implementation Details. The pre-trained ViT-L CLIP and R(2+1)D are used to extract 2D and 3D visual features, respectively, which are then concatenated to represent low-level visual features. The pre-trained HTSAT-RoBERTa-fusion CLAP is used to extract audio features. We adopt the standard HAN model (1-layer and 512-dim) in this task. AdamW optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay $= 1e-3$ is used to train the model. A learning rate scheduling of linear warm-up for 10 epochs to the peak learning rate of $3e-4$ and cosine annealing decay to the minimum learning rate of $3e-6$ is adopted. The batch size and the number of total training epochs are 16 and 120, respectively. We clip the gradient norm at 1.0 during training.

Quantitative Results. From Table 5.6, we observe that our baseline method performs on par with the previous state-of-the-art method CMBS [38]. When our method is applied to the model, the accuracy leaps from 75.3 to 80.4, indicating the generalizability of our method. In addition, we surpass CMBS [38] and have become the new state-of-the-art on the weakly-supervised AVE task with an improvement of **4.4** in accuracy.

5. Experiments



Method	Accuracy(%)
VGG-like, VGG-19 features	
AVEL [32]	66.7
AVSDN [33]	67.3
CMAN [93]	70.4
AVRB [94]	68.9
AVIN [95]	69.4
AVT [34]	70.2
CMRAN [36]	72.9
PSP [37]	73.5
CMBS [38]	74.2
VGG-like, Res-151 features	
AVEL [32]	71.6
AVSDN [33]	74.2
CMRAN [36]	75.3
CMBS [38]	76.0
CLAP, CLIP, R(2+1)D features	
HAN	75.3
VALOR	80.4

Table 5.6: **Results on the AVE task.**



Chapter 6

Conclusion

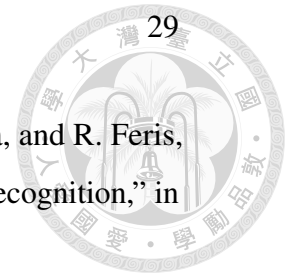
We propose **Visual-Audio Label Elaboration (VALOR)** for weakly-supervised Audio-Visual Video Parsing. By harnessing large-scale pre-trained contrastive models CLIP and CLAP, we generate fine-grained temporal labels in audio and visual modalities, providing explicit supervision to guide the learning of AVVP models. We show that utilizing modality-independent pre-trained models (CLIP and CLAP) and generating modality-aware labels are essential for AVVP. VALOR outperforms all the previous works when comparing in a fair setting, demonstrating its effectiveness. In addition, we demonstrate the generalizability of our method in the Audio-Visual Event Localization task, where we improve the baseline greatly and achieve a state-of-the-art result.

Limitations. While our method performs well on the AVVP task, it is uncertain whether it will maintain this efficacy when the number of events to classify expands. Moreover, because CLIP is not adept at capturing fine-grained visual information, it may fail to generate precise labels when the subject of the event to be identified is small or when the video quality is poor, potentially confounding the model.

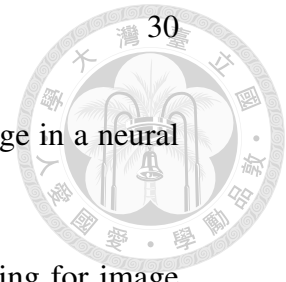


Reference

- [1] Y. Tian, D. Li, and C. Xu, “Unified multisensory perception: Weakly-supervised audio-visual video parsing,” in *ECCV*. Springer, 2020, pp. 436–454. [iv](#), [v](#), [vii](#), [1](#), [2](#), [5](#), [6](#), [16](#), [20](#), [22](#)
- [2] H. Cheng, Z. Liu, H. Zhou, C. Qian, W. Wu, and L. Wang, “Joint-modal label denoising for weakly-supervised audio-visual video parsing,” in *ECCV*. Springer, 2022, pp. 431–448. [v](#), [11](#), [16](#), [22](#)
- [3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018. [1](#), [12](#)
- [4] Q. Song, B. Sun, and S. Li, “Multimodal sparse transformer network for audio-visual speech recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022. [1](#), [12](#)
- [5] B. Shi, W.-N. Hsu, and A. Mohamed, “Robust self-supervised audio-visual speech recognition,” in *INTERSPEECH*, 2022. [1](#), [12](#)
- [6] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, “Audiovisual slowfast networks for video recognition,” *arXiv preprint arXiv:2001.08740*, 2020. [1](#), [12](#)
- [7] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, “Listen to look: Action recognition by previewing audio,” in *CVPR*, 2020, pp. 10 457–10 467. [1](#), [12](#)



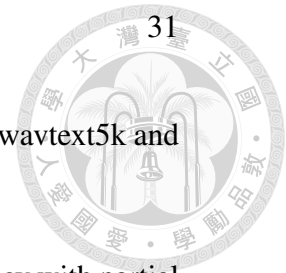
- [8] R. Panda, C.-F. R. Chen, Q. Fan, X. Sun, K. Saenko, A. Oliva, and R. Feris, “Adamml: Adaptive multi-modal learning for efficient video recognition,” in *ICCV*, 2021, pp. 7576–7585. 1, 12
- [9] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, “Foley music: Learning to generate music from videos,” in *ECCV*. Springer, 2020, pp. 758–775. 1, 12
- [10] K. Su, X. Liu, and E. Shlizerman, “Audeo: Audio generation for a silent performance video,” *NeurIPS*, vol. 33, pp. 3325–3337, 2020. 1, 12
- [11] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, “Visually informed binaural audio generation without binaural audios,” in *CVPR*, 2021, pp. 15 485–15 494. 1, 12
- [12] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, “Pano-avqa: Grounded audio-visual question answering on 360deg videos,” in *ICCV*, 2021, pp. 2031–2041. 1, 13
- [13] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, “Learning to answer questions in dynamic audio-visual scenarios,” in *CVPR*, 2022, pp. 19 108–19 118. 1, 13
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*. PMLR, 2021, pp. 8748–8763. 3, 7, 8
- [15] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP*. IEEE, 2023, pp. 1–5. 3, 8, 15

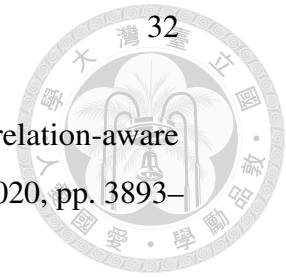


- [16] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015. 3, 9
- [17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778. 5
- [18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *CVPR*, 2018, pp. 6450–6459. 5
- [19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, “Cnn architectures for large-scale audio classification,” in *ICASSP*. IEEE, 2017, pp. 131–135. 6
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *NeurIPS*, vol. 30, 2017. 6
- [21] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016. 6
- [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016, pp. 2818–2826. 6
- [23] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, “On metric learning for audio-text cross-modal retrieval,” in *INTERSPEECH*, 2022. 8
- [24] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “Audioclip: Extending clip to image, text and audio,” in *ICASSP*. IEEE, 2022, pp. 976–980. 8
- [25] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP*. IEEE, 2023, pp. 1–5. 8

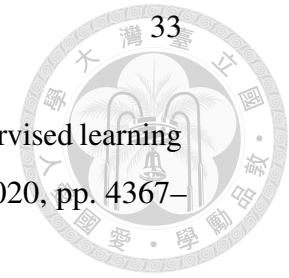
REFERENCE

- [26] S. Deshmukh, B. Elizalde, and H. Wang, “Audio retrieval with wavtext5k and clap training,” *arXiv preprint arXiv:2209.14275*, 2022. 8
- [27] Y.-B. Lin and Y.-C. F. Wang, “Exploiting audio-visual consistency with partial supervision for spatial audio generation,” in *AAAI*, vol. 35, no. 3, 2021, pp. 2056–2063. 11, 12
- [28] S. Mo and Y. Tian, “Multi-modal grouping network for weakly-supervised audio-visual video parsing,” in *NeurIPS*, 2022. 11, 16
- [29] J. Yu, Y. Cheng, R.-W. Zhao, R. Feng, and Y. Zhang, “Mm-pyramid: Multi-modal pyramid attentional network for audio-visual event localization and video parsing,” in *ACM MM*, 2022, pp. 6241–6249. 11, 16
- [30] Y. Wu and Y. Yang, “Exploring heterogeneous clues for weakly-supervised audio-visual video parsing,” in *CVPR*, 2021, pp. 1326–1335. 11, 16
- [31] J. Zhou, D. Guo, Y. Zhong, and M. Wang, “Improving audio-visual video parsing with pseudo visual labels,” *arXiv preprint arXiv:2303.02344*, 2023. 11, 16
- [32] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, “Audio-visual event localization in unconstrained videos,” in *ECCV*, 2018, pp. 247–263. 12, 16, 24, 25, 26
- [33] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, “Dual-modality seq2seq network for audio-visual event localization,” in *ICASSP*. IEEE, 2019, pp. 2002–2006. 12, 16, 26
- [34] Y.-B. Lin and Y.-C. F. Wang, “Audiovisual transformer with instance attention for audio-visual event localization,” in *ACCV*, 2020. 12, 26
- [35] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, “Dual attention matching for audio-visual event localization,” in *ICCV*, 2019, pp. 6292–6300. 12





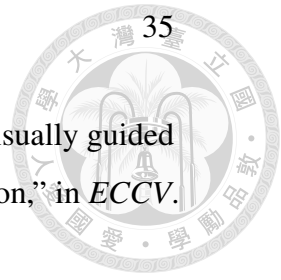
- [36] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, “Cross-modal relation-aware networks for audio-visual event localization,” in *ACM MM*, 2020, pp. 3893–3901. 12, 26
- [37] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, “Positive sample propagation along the audio-visual event line,” in *CVPR*, 2021, pp. 8436–8444. 12, 26
- [38] Y. Xia and Z. Zhao, “Cross-modal background suppression for audio-visual event localization,” in *CVPR*, 2022, pp. 19 989–19 998. 12, 25, 26
- [39] D. Hu, X. Li *et al.*, “Temporal multimodal learning in audiovisual speech recognition,” in *CVPR*, 2016, pp. 3574–3582. 12
- [40] L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, “A multi-view approach to audio-visual speaker verification,” in *ICASSP*. IEEE, 2021, pp. 6194–6198. 12
- [41] S. Shon, T.-H. Oh, and J. Glass, “Noise-tolerant audio-visual online person verification using an attention-based neural network fusion,” in *ICASSP*. IEEE, 2019, pp. 3995–3999. 12
- [42] Y. Qian, Z. Chen, and S. Wang, “Audio-visual deep neural network for robust person verification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021. 12
- [43] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, “Audio-visual person recognition in multimedia data from the iarpa janus program,” in *ICASSP*. IEEE, 2018, pp. 3031–3035. 12
- [44] J. S. Chung, B.-J. Lee, and I. Han, “Who said that?: Audio-visual speaker diarisation of real-world meetings,” in *INTERSPEECH*, 2019. 12



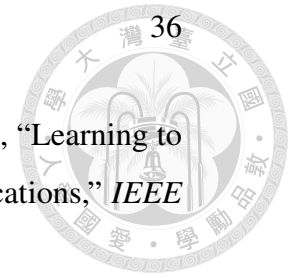
- [45] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, “Self-supervised learning for audio-visual speaker diarization,” in *ICASSP*. IEEE, 2020, pp. 4367–4371. 12
- [46] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, “Epic-fusion: Audio-visual temporal binding for egocentric action recognition,” in *ICCV*, 2019, pp. 5492–5501. 12
- [47] B. Korbar, D. Tran, and L. Torresani, “Scsampler: Sampling salient clips from video for efficient action recognition,” in *ICCV*, 2019, pp. 6232–6242. 12
- [48] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, “On training targets and objective functions for deep-learning-based audio-visual speech enhancement,” in *ICASSP*. IEEE, 2019, pp. 8077–8081. 12
- [49] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” in *INTERSPEECH*, 2018. 12
- [50] —, “My lips are concealed: Audio-visual speech enhancement through obstructions,” in *INTERSPEECH*, 2019. 12
- [51] M. Sadeghi and X. Alameda-Pineda, “Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders,” in *ICASSP*. IEEE, 2020, pp. 7534–7538. 12
- [52] J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn, “Looking into your speech: Learning cross-modal affinity for audio-visual speech separation,” in *CVPR*, 2021, pp. 1336–1345. 12
- [53] Z. Kang, M. Sadeghi, R. Horaud, X. Alameda-Pineda, J. Donley, and A. Kumar, “The impact of removing head movements on audio-visual speech enhancement,” in *ICASSP*. IEEE, 2022, pp. 7302–7306. 12



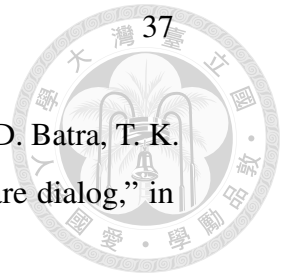
- [54] R. Gao, R. Feris, and K. Grauman, “Learning to separate object sounds by watching unlabeled video,” in *ECCV*, 2018, pp. 35–53. 12
- [55] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in *ECCV*, 2018, pp. 570–586. 12
- [56] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, “Self-supervised audio-visual co-segmentation,” in *ICASSP*. IEEE, 2019, pp. 2357–2361. 12
- [57] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, “The sound of motions,” in *ICCV*, 2019, pp. 1735–1744. 12
- [58] X. Xu, B. Dai, and D. Lin, “Recursive visual sound separation using minus-plus net,” in *ICCV*, 2019, pp. 882–891. 12
- [59] R. Gao and K. Grauman, “Co-separating sounds of visual objects,” in *ICCV*, 2019, pp. 3879–3888. 12
- [60] M. Chatterjee, J. Le Roux, N. Ahuja, and A. Cherian, “Visual scene graphs for audio source separation,” in *ICCV*, 2021, pp. 1204–1213. 12
- [61] Y. Tian, D. Hu, and C. Xu, “Cyclic co-learning of sounding object visual grounding and sound separation,” in *CVPR*, 2021, pp. 2745–2754. 12
- [62] V. K. Kurmi, V. Bajaj, B. N. Patro, K. Venkatesh, V. P. Namboodiri, and P. Jyothi, “Collaborative learning to generate audio-video jointly,” in *ICASSP*. IEEE, 2021, pp. 4180–4184. 12
- [63] R. Gao and K. Grauman, “2.5 d visual sound,” in *CVPR*, 2019, pp. 324–333. 12
- [64] K. K. Parida, S. Srivastava, and G. Sharma, “Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention,” in *WACV*, 2022, pp. 3347–3356. 12



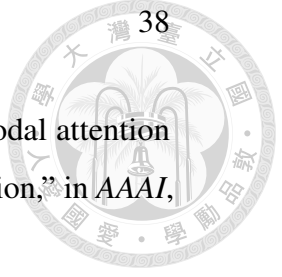
- [65] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu, “Sep-stereo: Visually guided stereophonic audio generation by associating source separation,” in *ECCV*. Springer, 2020, pp. 52–69. 12
- [66] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” in *ICCV*, 2019, pp. 9459–9468. 12
- [67] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, “Pose-controllable talking face generation by implicitly modularized audio-visual representation,” in *CVPR*, 2021, pp. 4176–4186. 12
- [68] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan, and Y. Yang, “Seeg: Semantic energized co-speech gesture generation,” in *CVPR*, 2022, pp. 10 473–10 482. 12
- [69] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, “Dancing to music,” *NeurIPS*, vol. 32, 2019. 12
- [70] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, “Ai choreographer: Music conditioned 3d dance generation with aist++,” in *ICCV*, 2021, pp. 13 401–13 412. 12
- [71] X. Li, D. Hu, and X. Lu, “Image2song: Song retrieval via bridging image content and lyric words,” in *ICCV*, 2017, pp. 5649–5658. 12
- [72] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto, “Cross-modal embeddings for video and audio retrieval,” in *ECCV workshops*, 2018, pp. 0–0. 12
- [73] R. Arandjelovic and A. Zisserman, “Objects that sound,” in *ECCV*, 2018, pp. 435–451. 12
- [74] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in *ECCV*, 2018, pp. 631–648. 12



- [75] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, “Learning to localize sound sources in visual scenes: Analysis and applications,” *IEEE TPAMI*, vol. 43, no. 5, pp. 1605–1619, 2019. 12
- [76] X. Hu, Z. Chen, and A. Owens, “Mix and localize: Localizing sound sources in mixtures,” in *CVPR*, 2022, pp. 10 483–10 492. 12
- [77] D. Hu, Y. Wei, R. Qian, W. Lin, R. Song, and J.-R. Wen, “Class-aware sounding objects localization via audiovisual correspondence,” *IEEE TPAMI*, vol. 44, no. 12, pp. 9844–9859, 2021. 12
- [78] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, “Soundspaces: Audio-visual navigation in 3d environments,” in *ECCV*. Springer, 2020, pp. 17–36. 13
- [79] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, “Look, listen, and act: Towards audio-visual embodied navigation,” in *ICRA*. IEEE, 2020, pp. 9701–9707. 13
- [80] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, “Learning to set waypoints for audio-visual navigation,” in *ICLR*, 2021. 13
- [81] C. Chen, Z. Al-Halah, and K. Grauman, “Semantic audio-visual navigation,” in *CVPR*, 2021, pp. 15 516–15 525. 13
- [82] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, “Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds,” *IEEE Robotics and Automation Letters*, 2023. 13
- [83] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu, “Sound adversarial audio-visual navigation,” in *ICLR*, 2022. 13
- [84] S. Majumder, Z. Al-Halah, and K. Grauman, “Move2hear: Active audio-visual source separation,” in *ICCV*, 2021, pp. 275–285. 13



- [85] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson *et al.*, “Audio visual scene-aware dialog,” in *CVPR*, 2019, pp. 7558–7567. 13
- [86] C. Hori, A. Cherian, T. K. Marks, and T. Hori, “Joint student-teacher learning for audio-visual scene-aware dialog,” in *INTERSPEECH*, 2019, pp. 1886–1890. 13
- [87] I. Schwartz, A. G. Schwing, and T. Hazan, “A simple baseline for audio-visual scene-aware dialog,” in *CVPR*, 2019, pp. 12 548–12 558. 13
- [88] S. Geng, P. Gao, M. Chatterjee, C. Hori, J. Le Roux, Y. Zhang, H. Li, and A. Cherian, “Dynamic graph representation learning for video dialog via multi-modal shuffled transformers,” in *AAAI*, vol. 35, no. 2, 2021, pp. 1415–1423. 13
- [89] A. Shah, S. Geng, P. Gao, A. Cherian, T. Hori, T. K. Marks, J. Le Roux, and C. Hori, “Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning,” in *ICASSP*. IEEE, 2022, pp. 7732–7736. 13
- [90] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang, “Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing,” *NeurIPS*, vol. 34, pp. 11 449–11 461, 2021. 16
- [91] X. Jiang, X. Xu, Z. Chen, J. Zhang, J. Song, F. Shen, H. Lu, and H. T. Shen, “Dhnn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing,” in *ACM MM*, 2022, pp. 719–727. 16
- [92] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *ICASSP*. IEEE, 2017, pp. 776–780. 24

- 
- [93] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, “Cross-modal attention network for temporal inconsistent audio-visual event localization,” in *AAAI*, vol. 34, 2020, pp. 279–286. 26
- [94] J. Ramaswamy and S. Das, “See the sound, hear the pixels,” in *WACV*, 2020, pp. 2970–2979. 26
- [95] J. Ramaswamy, “What makes the sound?: A dual-modality interacting network for audio-visual event localization,” in *ICASSP*. IEEE, 2020, pp. 4372–4376. 26