UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing

Yung-Hsuan Lai[1,†], Janek Ebbers[3], Yu-Chiang Frank Wang[1,2], François Germain[3], Michael Jeffrey Jones[3], Moitreya Chatterjee[3,‡],

[1] National Taiwan University [2] NVIDIA, Taiwan [3] Mitsubishi Electric Research Labs (MERL)

[†]r10942097@ntu.edu.tw [‡]metro.smiles@gmail.com

Abstract

Audio-Visual Video Parsing (AVVP) is a challenging task that aims to localize both uni-modal and multi-modal events in a weakly-supervised setting, where only modality-agnostic, video-level labels are available during training. While prior works seek to generate segment-level pseudo-labels to better guide model training, their performance is limited primarily by the lack of inter-segment dependencies in the label generation process and the general bias towards predicting events that are absent in a segment. To address these issues, we propose a novel approach called Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing (UWAV). Additionally, our innovative approach factors in the uncertainty associated with these estimated pseudo-labels and incorporates a feature mixup based training regularization for improved training. Empirical results show that UWAV outperforms state-of-the-art methods for the AVVP task on multiple metrics.

1. Introduction

Audio-visual learning has become a key focus in multi-modal research. Various audio-visual learning tasks have been widely explored, including audio-visual segmentation [14] and audio-visual event localization [3, 9]. However, many of these tasks assume perfect temporal alignment between audio and visual streams – an assumption that frequently fails in real-world scenarios.

In this work we explore the task of Audio-Visual Video Parsing (AVVP) [10], which aims to recognize all audio-only, visual-only, and audio-visual events occurring in each one-second segment of a video (see Fig. 1). From a machine learning standpoint, AVVP poses two key challenges: (i) the audio and visual events that occur may not be temporally aligned, *e.g.* a sound may be heard before its source appears on screen and (ii) due to the high cost of segment-level annotation, only modality-agnostic, video-level labels may be available during training. Recent work addresses these challenges by generating richer pseudo-labels, either at the video-level [1] or segment-level [3, 5]. Notably, VALOR [3] leverages large-scale pre-trained foundation models (*e.g.* CLIP [6] and CLAP [12]) along with ground-truth video-level labels to generate segment-level pseudo-labels for each of the two modalities. Audio/Visual segments (*e.g.* the audio corresponding to the segment in question and the visual frame at the center of the segment) are fed into CLAP/CLIP, one segment at a time, to generate these pseudo-labels. While promising, these methods are limited by the lack of a global video understanding during the pseudo-label generation process.

To address this oversight and other shortcomings in existing pseudo-label generation methods, we introduce UWAV (Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing), a novel framework that generates improved segment-level pseudo-labels for better training of the inference module. UWAV uses transformer modules [11] pre-trained on a large-scale, supervised audio-visual event localization dataset [2] to capture temporal dependencies across video segments. Subsequently, we use this pre-trained model to generate modality-specific, segment-level pseudo-labels on a target, small-scale dataset with only weak (video-level) supervision while factoring in the uncertainty associated with these pseudo-labels. Moreover, UWAV addresses the critical class imbalance issue in the pseudo-label enriched training data – where most event classes are absent (*i.e.*, they are negative) in any segment – by introducing a class-balanced loss re-weighting strategy to focus on learning positive events. These components together with a feature-mixup based training enable UWAV to outperform state-of-the-art approaches on the benchmark Look, Listen, and Parse dataset [10].

2. Related Works

To address the challenges of the AVVP task, [10] proposed a Hybrid Attention Network (HAN) and a learnable Multi-modal Multiple Instance Learning (MMIL) pooling module. The HAN model facilitates the exchange of information within and across modalities using self-attention and cross-attention layers, while the MMIL pooling module aggregates segment-level event probabilities from both modalities to produce video-level probabilities. Building on this foundation, recent works advanced the field from the following two perspectives. The first group of studies focuses on enhancing model architectures [15]. In particular, [4] proposed the Multi-modal Grouping Network (MGN) to explicitly group semantically similar features within each modality, while [13] proposed Multi-modal Pyramid Attentional Network (MM-Pyramid) to capture events of varying

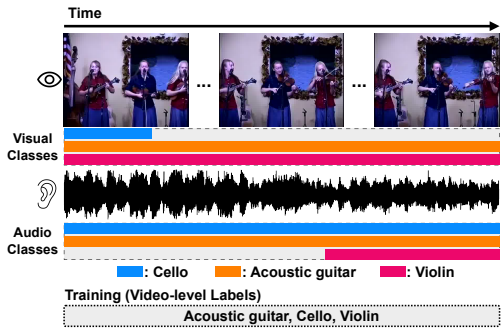

Figure 1: **A weakly-supervised AVVP task example.** Events, considered in this task, might be uni-modal or multi-modal. Even multimodal events, may not be temporally aligned in the audio and visual modalities, *e.g.* the cello might only be visible in the first few seconds but might produce music, throughout the video.

durations by extracting features at multiple temporal scales. Our proposed method is orthogonal to this line of research and can be integrated with any of these backbones.

The second direction focuses on generating pseudo-labels for improved training, either at the video-level [1] or at the segment-level [3, 5]. Such methods generally utilize the popular CLIP [6] and CLAP [12] features along with ground-truth video-level labels to predict pseudo-labels for each modality on a per-segment basis. PPL [5] uses the HAN model itself to generate pseudo-labels by constructing prototype features for each class and uses them for clustering, which however might not scale to smaller datasets. Differently, our method captures inter-segment dependencies for pseudo-label generation and scales to smaller datasets.

3. Preliminaries

Problem Formulation: The AVVP task aims to recognize all visible and/or audible events in each one-second segment of a video. Specifically, an audible video is split into T one-second segments, denoted as $\{V_t, A_t\}_{t=1}^{T}$. Each segment is annotated with a pair of ground-truth labels $y_t^v \in \{0,1\}^C, y_t^a \in \{0,1\}^C$, where y_t^v denotes visual events, y_t^a denotes audio events, and C denotes the total number of events in the pre-defined event set of the data. However, (y_t^v, y_t^a) are unavailable during training, instead only the modality-agnostic, video-level labels $y \in \{0,1\}^C$ are available, where 1 indicates the presence of an event at any time (either in the audio or visual stream or both) while 0 indicates an event's absence in the video.

Pseudo-Label Based AVVP Framework: The Hybrid Attention Network (HAN) [10] is a commonly used model for the AVVP task. First, pre-trained visual and audio backbones are used to extract features from the visual and audio segments, respectively, which are then projected to two d-dimensional feature spaces. The resulting visual segment-level and audio segment-level features are provided as input to the HAN model. In the model, information across segments within a modality and across modalities is exchanged through self-attention and cross-attention layers. Finally, a classifier, shared across both modalities, transforms the visual segment-level features (*resp.* audio segment-level features), obtained from the HAN model, into visual segment-level logits $\{z_t^v\}_{t=1}^{T} \in \mathbb{R}^{T\times C}$ (*resp.* audio segment-level logits $\{z_t^a\}_{t=1}^{T}$). Segment-level probabilities $\{p_t^v\}_{t=1}^{T}, \{p_t^a\}_{t=1}^{T} \in \mathbb{R}^{T\times C}$ are then obtained by applying the sigmoid function on $\{z_t^v\}_{t=1}^{T}$ and $\{z_t^a\}_{t=1}^{T}$.

Since only video-level labels y are available during training, an attentive MMIL pooling module [10] is introduced to predict video-level probabilities $p \in \mathbb{R}^C$:

$$p = \sum_{m=\{v,a\}} \sum_{t=1}^{T} W_t^m \odot p_t^m, \tag{1}$$

where $W_t^m \in \mathbb{R}^C$ is the weight output by the learnable MMIL pooling module for each modality and each segment, and $\odot$ denotes the element-wise product. The HAN model is then optimized with the binary cross-entropy (BCE) loss between the estimated video-level probabilities p and video-level labels y: $\mathcal{L}_{video} = \text{BCE}(p, y)$.

4. Proposed Approach

At a high level, UWAV generates better segment-level pseudo-labels to improve the training of the HAN model. Moreover, UWAV factors in the uncertainty associated with these pseudo-labels and addresses the imbalance in the training data. Figure 2 shows an overview of UWAV.

4.1. Uncertainty-aware Pseudo-Label Synthesis

One major issue that plagues prior pseudo-label generation-based works [3] is their inability to capture the full context of the video when generating pseudo-labels for individual segments. To plug this void, we pre-train transformer modules [11] (one for audio, one for video input) on a supervised, audio-visual event localization dataset, enabling them to produce segment-level predictions with awareness of the entire video context.

Pre-Training Pseudo-Label Generation Modules: Given an audible video of duration T' seconds from the pre-training dataset (e.g. the large-scale, supervised UnAV [2] dataset), we split the video into T' one-second segments $\{V'_t, A'_t\}_{t=1}^{T'}$, with corresponding audio-visual event labels $y_t^{av\prime} \in \{0,1\}^{C'}$, where 1 indicates the presence of an event in both modalities and 0 its absence in at least one modality, while C' denotes the total number of event classes in the pre-training dataset. Next, the video frame at the temporal center of the visual segment is transformed into d_1-D visual features $G_0^{v\prime} = \{g_{0,t}^{v\ \prime}\}_{t=1}^{T'} \in \mathbb{R}^{T'\times d_1}$ with CLIP's [6] image encoder. These features are then encoded via a transformer [11] with $L = 5$ encoder blocks. Concurrently, we convert each event category label in the pre-training dataset into a textual event feature $e_c^{CLIP\prime} \in \mathbb{R}^{d_1}$ by filling

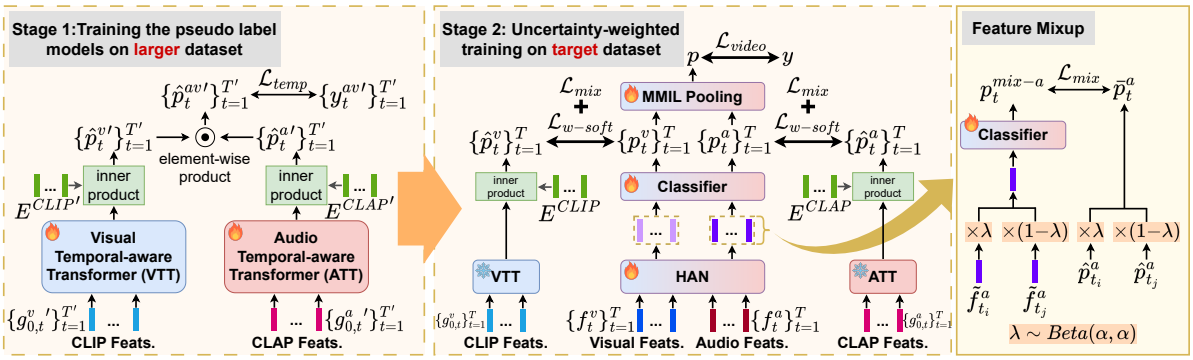

Figure 2: **An overview of our UWAV framework.**

Table 1: **Comparison with state-of-the-art methods on the LLP dataset.** Best performances are in **bold**.

Method	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
HAN [10]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
JoMoLD [1]	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
VALOR [3]	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
PPL [5]	**65.9**	66.7	61.9	64.8	63.7	57.3	64.3	54.3	59.9	**57.9**
CoLeaf [8]	64.2	64.4	59.3	62.6	62.5	57.6	63.2	54.2	57.9	55.6
LEAP [15]	62.7	65.6	59.3	62.5	61.8	56.4	63.1	54.1	57.8	55.0
UWAV (Ours)	64.2	**70.0**	**63.4**	**65.9**	**63.9**	**58.6**	**66.7**	**57.5**	**60.9**	57.4

Table 2: **Accuracy of the generated pseudo-labels.**

Method	Segment-level				
	A	V	AV	Type	Event
VALOR [3]	**80.5**	61.7	55.7	66.0	74.6
PPL [5]	61.7	61.8	57.5	60.6	59.4
UWAV (Ours)	78.4	**74.5**	**65.5**	**72.8**	**78.4**

in the pre-defined caption template: "A photo of <EVENT NAME>" with the corresponding event name and passing it to CLIP's text encoder. With the visual segment-level features $G_L^{v\prime} = \{g_{L,t}^{v}{}'\}_{t=1}^{T'} \in \mathbb{R}^{T' \times d_1}$ and the textual event features $E^{CLIP\prime} = \{e_c^{CLIP\prime}\}_{c=1}^{C'} \in \mathbb{R}^{C' \times d_1}$ in place, we derive visual segment-level logits $\hat{z}_t^{v\prime} \in \mathbb{R}^{C'}$ and probabilities $\hat{p}_t^{v\prime}$ as follows:

$$\hat{p}_t^{v\prime} = \text{Sigmoid}(\hat{z}_t^{v\prime}),\ \hat{z}_t^{v\prime} = E^{CLIP\prime} \cdot g_{L,t}^{v}{}'^{\top}. \quad (2)$$

Similar operations are performed in the audio pseudo-label generation pipeline. The raw waveforms corresponding to the 1-second audio segments are transformed into d_2-D audio features $G_0^{a\prime} \in \mathbb{R}^{T' \times d_2}$ with CLAP's [12] audio encoder and fed into the audio transformer. The textual event features $E^{CLAP\prime} \in \mathbb{R}^{C' \times d_2}$ are generated with the caption template: "This is the sound of <EVENT NAME>" by passing it through CLAP's text encoder. Audio segment-level logits $\hat{z}_t^{a\prime} \in \mathbb{R}^{C'}$ and probabilities $\hat{p}_t^{a\prime}$ can then be derived as: $\hat{p}_t^{a\prime} = \text{Sigmoid}(\hat{z}_t^{a\prime}),\ \hat{z}_t^{a\prime} = E^{CLAP\prime} \cdot g_t^{a\prime\top}$.

Since the events occurring in the pre-training dataset (UnAV) are audio-visual, we multiply the segment-level visual and audio event probabilities to obtain the event probabilities and train the transformers with BCE loss:

$$\mathcal{L}_{temp} = \text{BCE}(\hat{p}_t^{av\prime}, y_t^{av\prime}),\ \hat{p}_t^{av\prime} = \hat{p}_t^{v\prime} \odot \hat{p}_t^{a\prime}. \quad (3)$$

Uncertainty-weighted Pseudo-Label Training: With the pre-trained pseudo-label generation modules in place, we proceed to employ them on the target dataset for the AVVP task. To do so, one approach is to follow VALOR [3], to determine thresholds first and generate binary segment-level pseudo-labels for both modalities. However, the generated pseudo-labels could potentially be noisy, leading to occasionally incorrect training signals. To ameliorate this problem, we propose an uncertainty-weighted pseudo-label based training scheme by leveraging the confidence of the pseudo-label estimation module (associated with the predicted pseudo-label) to weigh the training signal for the inference module. This confidence score serves as a measure of the pseudo-label generation module's uncertainty of its prediction. This may be represented as:

$$\hat{p}_t^{v} = \text{Sigmoid}(\hat{z}_t^{v} - \theta^{v}) \odot y;\ \hat{p}_t^{a} = \text{Sigmoid}(\hat{z}_t^{a} - \theta^{a}) \odot y, \quad (4)$$

where $\hat{z}_t^{v}$ and $\hat{z}_t^{a}$ are segment-level visual and audio logits (generated in the same manner as before) on the target dataset, and $\theta^{v} \in \mathbb{R}^{C}$ and $\theta^{a} \in \mathbb{R}^{C}$ are the visual and audio event thresholds. With the uncertainty-weighted pseudo-labels in place, the inference module (HAN) can be trained with the following uncertainty-weighted pseudo-label loss:

$$\mathcal{L}_{soft} = \text{BCE}(p_t^{v}, \hat{p}_t^{v}) + \text{BCE}(p_t^{a}, \hat{p}_t^{a}). \quad (5)$$

4.2. Class-balanced Loss Re-weighting

Besides the aforementioned challenges of the AVVP task, most of the events in the event set are absent/negative

in any given segment of a video. As a result, the training of the inference module is dominated by the loss from the negative events while the positive ones tend to get ignored. To address this issue, we introduce a *class-balanced loss re-weighting* strategy to re-balance the uncertainty-weighted pseudo-label loss between the negative and positive events. Specifically, the loss from the positive events is multiplied by a weight proportional to the frequency of the segments with the negative events in the pseudo-labels and vice-versa:

$$\mathcal{L}_{w-soft} = \sum_{m\in\{v,a\}} w^m_{pos} \cdot y \cdot \text{BCE}(p^m_t, \hat{p}^m_t) + w^m_{neg} \cdot (1-y) \cdot \text{BCE}(p^m_t, \hat{p}^m_t), \quad (6)$$

$$w^m_{pos} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}(1-\hat{y}^m_{i,t,c})}{NTC} \times W, \quad (7)$$

$$w^m_{neg} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{y}^m_{i,t,c}}{NTC}, \quad (8)$$

where N denotes the number of videos in the training set, and W is a hyper-parameter set to 0.5.

For more improved training, we exploy a feature-mixup [7] loss ($\mathcal{L}_{mix}$) on the audio ($G^a_L{}'$) and visual ($G^v_L{}'$) features, as well. In summary, the inference module is trained on the AVVP task with the following losses:

$$\mathcal{L}_{total} = \mathcal{L}_{w-soft} + \mathcal{L}_{video} + \mathcal{L}_{mix}. \quad (9)$$

5. Experiments

5.1. Experimental Setup

Dataset: We evaluate all competing methods on the *Look, Listen, and Parse* (LLP) dataset [10] – the benchmark dataset for the AVVP task. The dataset consists of $11,849$ video clips sourced from YouTube. Each clip is 10 seconds long and represents one or more of 25 diverse event categories, such as human activities, animals, and musical instruments. We follow the official split [10] into training, validation, and testing sets, for our experiments.

Metrics: Following the official protocol [10], all models are evaluated using macro F1-scores calculated for the following settings: (i) audio-only (A), (ii) visual-only (V), (iii) audio-visual (AV), (iv) Type@AV (Type), and (v) Event@AV (Event). Type@AV is the mean of the A, V, and AV event F-scores, while Event@AV is the F1-score of all events regardless of the modality in which they occur. Evaluations are conducted at both segment- and event-levels.

5.2. Results

Comparison with Previous Methods: As shown in Table 1, UWAV surpasses previous methods across almost all metrics. Notably, we achieve a gain of 1.1% on segment-level Type@AV F-score and a 1% improvement on event-level Type@AV F-score, over our closest competitor PPL [5]. When compared to other recently published works, such as VALOR [3], CoLeaf [8] and LEAP [15], UWAV outperforms them by up to 3% on both segment and event-level Type@AV F-scores.

Accuracy of the Generated Pseudo-Labels: To evaluate the efficacy of our pseudo-label generation pipeline, we compare the accuracy of our generated pseudo-labels against those obtained from competing methods (with publicly available implementation) on the test set of the LLP dataset. As shown in Table 2, our pseudo-label generation scheme generates more accurate segment-level pseudo-labels compared to previous methods, by up to $\underline{6\%}$ on the segment-level Type@AV F-score.

6. Conclusions

In this work, we propose UWAV to incorporate inter-segment dependencies in the pseudo-label generation process for the AVVP task. In addition, by factoring in the uncertainty associated with these estimated pseudo-labels, correcting for class imbalance, and training with a feature mixup strategy, UWAV achieves state-of-the-art performance for the AVVP task.

References

[1] H. Cheng et al. Joint-modal label denoising for weakly-supervised audio-visual parsing. In *ECCV*, 2022. 1, 2, 3

[2] T. Geng et al. Dense-localizing audio-visual events in untrimmed videos. In *CVPR*, 2023. 1, 2

[3] Y.-H. Lai et al. Modality-independent teachers meet weakly-supervised audio-visual event parser. In *NeurIPS*, 2023. 1, 2, 3, 4

[4] S. Mo and Y. Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *NeurIPS*, 2022. 1

[5] K. K. Rachavarapu et al. Weakly-supervised audio-visual video parsing with prototype-based pseudo-labeling. In *CVPR*, 2024. 1, 2, 3, 4

[6] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[7] S. Ren et al. A simple data mixing prior for improving self-supervised learning. In *CVPR*, 2022. 4

[8] F. Sardari et al. Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing. In *ECCV*, 2024. 3, 4

[9] Y. Tian et al. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1

[10] Y. Tian et al. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 2, 3, 4

[11] A. Vaswani et al. Attention is all you need. In *NeurIPS*, 2017. 1, 2

[12] Y. Wu et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 1, 2, 3

[13] J. Yu et al. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM MM*, 2022. 1

[14] J. Zhou et al. Audio–visual segmentation. In *ECCV*, 2022. 1

[15] J. Zhou et al. Label-anticipated event disentanglement for audio-visual video parsing. *arXiv preprint arXiv:2407.08126*, 2024. 1, 3, 4