TEACHER-GUIDED PSEUDO SUPERVISION AND CROSS-MODAL ALIGNMENT FOR AUDIO-VISUAL VIDEO PARSING

Yaru Chen[1], *Ruohao Guo*[2], *Liting Gao*[1], *Yang Xiang*[1], *Qingyu Luo*[1], *Zhenbo Li*[3], *Wenwu Wang*[1]

[1]Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom
[2]School of Intelligence Science and Technology, Peking University, China
[3]College of Information and Electrical Engineering, China Agricultural University, China

ABSTRACT

Weakly-supervised audio-visual video parsing (AVVP) seeks to detect audible, visible, and audio-visual events without temporal annotations. Previous work has emphasized refining global predictions through contrastive or collaborative learning, but neglected stable segment-level supervision and class-aware cross-modal alignment. To address this, we propose two strategies: (1) an exponential moving average (EMA)-guided pseudo supervision framework that generates reliable segment-level masks via adaptive thresholds or top-k selection, offering stable temporal guidance beyond video-level labels; and (2) a class-aware cross-modal agreement (CMA) loss that aligns audio and visual embeddings at reliable segment–class pairs, ensuring consistency across modalities while preserving temporal structure. Evaluations on LLP and UnAV-100 datasets shows that our method achieves state-of-the-art (SOTA) performance across multiple metrics.

***Index Terms*—** Audio-visual video parsing, Weakly-supervised learning, Exponential moving average, Cross-modal agreement, Audio-visual learning

1 Introduction

Weakly-supervised audio-visual video parsing (AVVP) [1] aims to localize audible, visible, and audio-visual events in unconstrained videos. As illustrated in Fig. 1, only video-level annotations are available during training, making it highly challenging to infer precise temporal boundaries and modality-specific events. This task is critical for applications such as audio-visual understanding, event detection, and segmentation [2, 3, 4]. Recent AVVP studies have explored strategies such as multi-instance learning for weak labels [1, 5], attention mechanisms to highlight informative segments [6, 7], and contrastive or collaborative learning to exploit audio–visual correlations [8, 9]. Although these approaches improve performance, two key challenges remain.

First, the lack of segment-level supervision, due to training with only video-level labels, makes it difficult to achieve stable learning. Previous works often propagate coarse labels to all segments or apply simple thresholding, which introduces noise [10, 11]. Although large pre-trained models (e.g., CLIP [12], CLAP [13]) have been used to generate pseudo labels [14, 15], these are typically static and cannot be refined during training, leaving them prone to noise and domain mismatch. Thus, more reliable, dynamically updated pseudo supervision is needed. Second, most cross-modal approaches align modalities by maximizing global audio–visual similarity [8, 16], overlooking that different classes may occur in different modalities at different times. Without class-aware, segment-level alignment, models risk forcing mismatches between unrelated events, leading to suboptimal localization.

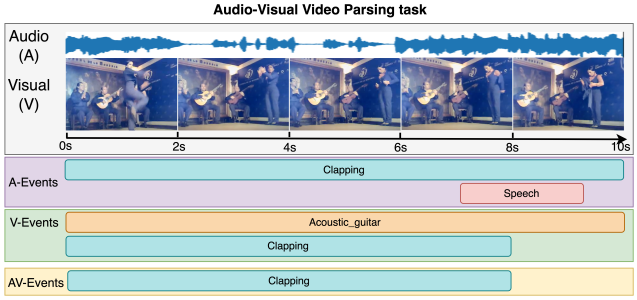

Fig. 1. Illustration of the AVVP task.

To overcome these limitations, we proposed E-CMA, which includes two strategies. (1) Exponential moving average (EMA)-guided pseudo supervision [17]. We adopt a teacher–student framework where the teacher, updated via EMA of student parameters, periodically generates segment-level pseudo masks from frame-wise predictions using adaptive thresholds or top-k selection per class. This transforms noisy video-level labels into stable, dynamically updated supervision, reducing error propagation from static pseudo labels. (2) Class-aware cross-modal agreement (CMA) loss. Rather than enforcing global similarity between audio and visual embeddings, CMA aligns modalities only at confident temporal–class positions where both modalities strongly indicate the same event. This selective alignment prevents over-matching asynchronous content and provides fine-grained, event-consistent supervision.

We evaluated our method on the widely used AVVP

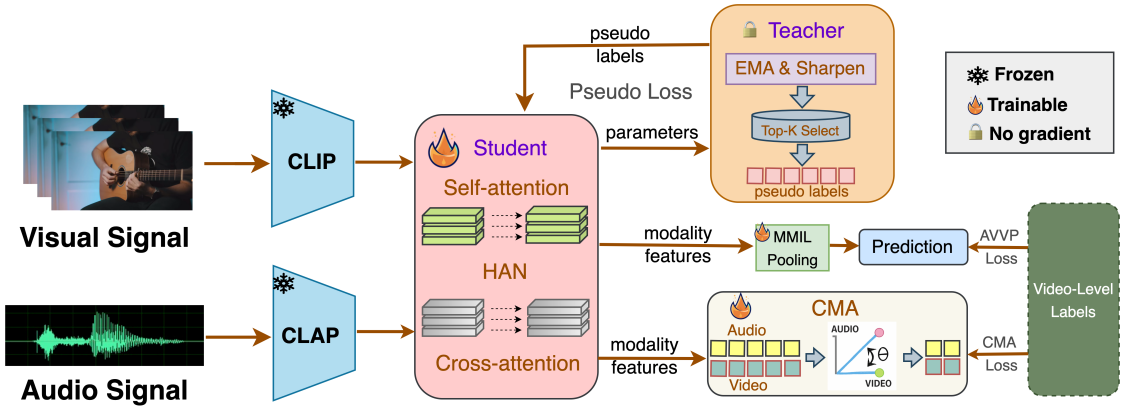

Fig. 2. Framework of E-CMA. The EMA Teacher provides stable pseudo labels to guide the Student, and the CMA module enforces class-aware cross-modal alignment. The final loss combines Pseudo loss, AVVP loss, and CMA loss.

benchmarks, the LLP dataset [1] and also the weakly-labeled UnAV-100 dataset [18]. Experimental results show that our approach achieves state-of-the-art (SOTA) performance on multiple metrics, which highlights the importance of stable pseudo supervision and fine-grained cross-modal agreement for advancing weakly-supervised AVVP.

2 Proposed Methods

2.1 Problem Statement

The goal of AVVP is to determine whether the event occurring in each time segment of a video is audio-only, video-only, or audio-visual event and to determine its category. Given a T second video $\mathcal{X}$ that is partitioned into T non-overlapping segments, each one second long. We denote the sequence as $\mathcal{X} = \{(x_t^a, x_t^v)\}_{t=1}^T$, where $x_t^a, x_t^v \in \mathbb{R}^d$ represent the audio and visual features at time t and d represents the dimension of the features. For each segment, we define the event labels as $y_t^a \in \{0,1\}^C, y_t^v \in \{0,1\}^C, y_t^{av} \in \{0,1\}^C$, where C is the number of event classes. The audio-visual event occurs only when both modalities detect the same event at time t, that means $y_t^{av} = y_t^a \odot y_t^v$, where $\odot$ denotes element-wise multiplication. AVVP is often a weakly supervised task, as only video-level labels are available during training, i.e. $y \in \{0,1\}^C$. In contrast, evaluation is performed with segment-level annotations, where fine-grained event boundaries are provided for both modalities.

2.2 Framework

As shown in Fig. 2, our framework builds upon the CoLeaF [8] baseline. We first extract audio and visual features using pre-trained CLAP [13] and CLIP [12] encoders. These features are refined by a Hierarchical Attention Network (HAN) [1] with self- and cross-attention to capture intra- and inter-modal dependencies. The updated segment representations are aggregated by Multimodal Multi-Instance Learning (MMIL) pooling [1] for video-level predictions. To enhance temporal supervision, we further introduce an EMA-guided teacher–student scheme. The teacher, updated by the exponential moving average, periodically generates stable segment-level pseudo masks that supplement weak video-level labels. In addition, a class-aware CMA loss is applied to enforce alignment between audio and visual embeddings in confident event segments. These novel designs enable more reliable segment-level learning and fine-grained cross-modal alignment.

2.3 Teacher-Guided Pseudo Supervision

To address the lack of segment-level annotations and the limitations of fixed, non-adaptive pseudo labels, we introduce an EMA-guided teacher–student framework, where a slowly updated teacher generates reliable segment-level masks, which are then used to supervise the training of the student network.

As shown in Fig. 2, the student network follows the architecture we introduced in Section 2.2, which includes CLAP and CLIP encoders, HAN aggregation, and MMIL pooling. It is updated by backpropagation with standard weakly-supervised objectives. The student network produces audio and visual probabilities $\hat{P}_t^a, \hat{P}_t^v \in [0,1]^C$. Then we fuse them into a joint prediction vector,

$$\hat{P}_t = \frac{1}{2}(\hat{P}_t^a + \hat{P}_t^v), \hat{P}_t \in [0,1]^C \quad (1)$$

The teacher network shares the same backbone architecture as the student network but is not optimized by gradient descent. Instead, its parameters are updated as an EMA [17] of the student's parameters, which maintains a weighted average of the previous student parameters to form a more stable teacher model. Let θ denote the parameters of the student at iteration k, and θ' denote the teacher parameters, thus:

$$\theta'_k = \alpha\theta'_{k-1} + (1-\alpha)\theta_k \quad (2)$$

where $\alpha \in [0,1)$ controls the update rate of the EMA teacher. This update ensures that the teacher evolves smoothly, making its predictions more stable than the student network.

To provide effective temporal supervision for the student, we transform the teacher's segment-level predictions into binary pseudo masks that indicate reliable event occurrences. The teacher's audio and visual predictions are first fused into a joint score:

$$\tilde{P}_t = \frac{1}{2}(\tilde{P}_t^a + \tilde{P}_t^v), \tilde{P}_t \in [0,1]^C \tag{3}$$

From $\tilde{P}_t$, we can derive binary pseudo masks $M_t \in \{0,1\}^C$ in two ways: adaptive thresholding or top-k selection. For adaptive thresholding, the threshold is dynamically adjusted by the mean prediction confidence for each class c, that means:

$$\tau = \gamma \cdot \frac{1}{T}\sum_{t=1}^{T} \tilde{P}_{t,c} \tag{4}$$

where γ is the scaling factor, then we define the pseudo mask:

$$M_{t,c} = \begin{cases} 1, & \text{if } \tilde{P}_{t,c} \geq \tau_c, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathbf{M} \in \{0,1\}^{T\times C}$ is the pseudo mask matrix of elements $M_{t,c}$, and $\tilde{P}_{t,c}$ indicate the teacher's predicted confidence score for class c at segment t. Alternatively, we can use $\mathrm{Top}-k$ selection to generate pseudo mask: if $t \in \mathrm{Top}{-}k\Big(\{\tilde{P}_{t,c}\}_{t=1}^{T}, k\Big)$, where $\mathrm{Top}{-}k$ returns the indices of the k highest confidence scores among all T segments for class c, and k is a hyperparameter which is the number of selected segments. After generating $\mathbf{M}$, we integrate them into the learning objective by enforcing consistency between the student's predictions and the pseudo labels, which we use a masked binary cross-entropy loss:

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{\|\mathbf{M}\|_1}\sum_{t=1}^{T}\sum_{c=1}^{C} \mathbf{M}_{t,c}\,\ell\Big(\hat{P}_{t,c}, 1\Big), \tag{6}$$

where $\|\mathbf{M}\|_1$ is the L1 norm. By this design, only the trusted segment–class pairs indicated by $\mathbf{M}$ contribute to the loss, while the remaining uncertain positions are ignored. This prevents noise accumulation and provides consistent temporal guidance beyond video-level labels.

2.4 Class-Aware Cross-Modal Alignment Loss

Although pseudo supervision provides reliable temporal masks, it does not explicitly enforce feature-level alignment across modalities. Hence, we introduce a class-aware CMA loss, which selectively encourages audio and visual embeddings to be consistent at confident segment–class positions.

Concretely, for each time step $t \in \{1,2,...,T\}$ and event class $c \in \{1,2,...,C\}$, we select valid segment-class pairs (t,c) which meet two conditions: (1) The predicted probabilities for both modalities $\hat{P}_{t,c}^a$ and $\hat{P}_{t,c}^v$ over their respective confidence thresholds τ_a and τ_v; (2) The video-level label $y_c = 1$, indicating the event c occurs in the video. We denote Ω as the set of these valid pairs, and apply the CMA loss for these pairs. For each pair $(t,c) \in \Omega$, we calculate the cosine similarity between audio and visual features:

$$s_{t,c} = \frac{(x_t^a)^\top x_t^v}{\|x_t^a\|_2 \cdot \|x_t^v\|_2} \tag{7}$$

Then, the CMA loss is formulated as the average cosine distance across all valid pairs:

$$\mathcal{L}_{\text{CMA}} = \frac{1}{|\Omega|}\sum_{(t,c)\in\Omega}(1 - s_{t,c}) \tag{8}$$

By restricting the loss to confident and label-consistent segment–class pairs, CMA suppresses noisy interactions and reinforces semantically meaningful alignment across modalities. Overall, the total loss l for our framework is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{AVVP}} + \mathcal{L}_{\text{pseudo}} + \mathcal{L}_{\text{CMA}} \tag{9}$$

Here, $\mathcal{L}_{\text{AVVP}}$ is the standard binary cross-entropy loss between predictions and ground-truth labels.

3 Experimental Results

3.1 Experimental Setup

Dataset. We evaluate our model on the LLP dataset [1], which is the benchmark for AVVP. It includes 11,849 10-seconds videos covering 25 event categories. We also perform tests on UnAV-100 [18], a large-scale dataset for audio-visual event localization, containing 10,790 videos and over 30k event instances from 100 classes. Following CoLeaF [8], we use only video-level labels for training UnAV-100.

Implementation Details. For the LLP dataset, we extract 768-dimensional audio and visual features using the pre-trained CLAP [13] and CLIP [12]. For the UnAV-100 dataset, we extract 2048-dimensional visual features using a two-stream I3D model (RGB + RAFT) [19], and 128-dimensional audio features using pre-trained VGGish [20].

Evaluation Metrics. We evaluate our model using F1-scores for three event types: audio (A), visual (V), and audio-visual (AV). A prediction is considered correct if the intersection-over-union (IoU) with ground truth exceeds 0.5. We computed the scores at the segment and event levels. The segment-level evaluation compares predictions and labels frame-by-frame, while event-level evaluation merges consecutive positive segments into a single event. We further report Type@AV, the average over A, V, and AV events, and Event@AV, which assesses the overall audio-visual event detection performance in each video.

3.2 Overall Performance Analysis

Table 1 and 2 show the comparative experimental results on the LLP and UnAV-100 datasets between our approach and previous SOTA methods. The results indicate that our method delivers superior performance on several key metrics, with particularly notable improvements in segment-level parsing, substantially outperforming existing approaches. Since our method does not incorporate text embeddings with audio and

Table 1. The performance of E-CMA and comparative methods in AVVP, with the best results highlighted in **bold** and the second results highlighted in text.

Model	Venue	Segment-level (%)					Event-level (%)				
		A	V	AV	Type@AV	Event@AV	A	V	AV	Type@AV	Event@AV
HAN [1]	ECCV'20	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MGN [16]	NeurIPS'22	60.8	55.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.0
MA [5]	CVPR'21	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
CMPAE [11]	CVPR'23	64.2	66.2	59.2	63.3	62.8	56.6	63.7	51.8	57.4	55.7
VALOR [14]	NeurIPS'23	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
CoLeaF [8]	ECCV'24	64.2	67.1	59.8	63.8	61.9	57.1	64.8	52.8	58.2	55.5
PPL [15]	CVPR'24	65.9	66.7	**61.9**	64.8	63.7	**57.3**	64.3	**54.3**	**59.9**	**57.9**
RLLD [21]	CVM'25	62.2	66.7	59.3	62.7	62.4	55.7	63.1	53.7	57.5	54.9
PPAE [9]	TPAMI'25	64.3	66.6	59.6	63.5	63.0	57.0	64.1	52.5	57.9	56.1
E-CMA	-	**66.1** (+0.2)	**69.9** (+2.8)	61.7	**65.9** (+1.1)	**65.4** (+1.7)	54.5	**66.6** (+1.8)	53.5	58.2	54.3

Table 2. Comparison of E-CMA performance on the weakly-labeled UnAV-100 dataset.

Method	**AV (Seg)**	**AV (Evn)**
HAN [1]	35.0	41.4
MA [5]	37.9	44.8
JoMoLD [10]	36.4	41.2
CMPAE [11]	39.7	43.8
CoLeaF [8]	41.5	**47.8**
E-CMA	**41.8** (+0.3)	47.4

visual features, we restrict the comparison to approaches that also do not employ such fusion.

As shown in Table 1, E-CMA achieves the best overall performance on the LLP dataset. At the segment level, it consistently outperforms previous approaches, reaching 66.1% and 69.9% on the audio and visual modalities, both new SOTA results. It also yields improvements on joint AV metrics, with 61.7% on AV and 65.4% on Event@AV. At the event level, E-CMA attains the highest visual score of 66.6% and competitive results across other metrics, including 58.2% on Type@AV and 54.3% on Event@AV. These results demonstrate that E-CMA not only enhances unimodal performance but also achieves more consistent cross-modal event parsing. On the UnAV-100 dataset with weakly supervised labels, E-CMA achieves 41.8% on AV (Seg), surpassing CoLeaF by +0.3%. For AV (Event), our model obtains 47.4%, which is competitive with the best baseline (47.8%). These results show that E-CMA not only enhances segment-level localization but remains strong event-level parsing ability under weak supervision.

3.3 Ablation Study

To assess the impact of the EMA and CMA modules, we conducted ablation study on the LLP dataset by removing them from our method. For fairness, we train the CoLeaF with the same feature extractor as our method, denoted as CoLeaF†.

As shown in Table 3, both modules contribute to the effectiveness of our framework. Specifically, removing CMA leads to drops in visual and audio-visual metrics at both the segment and event levels, indicating that CMA is crucial for enhancing cross-modal alignment. On the other hand, excluding EMA mainly affects event-level results, with a decrease in Event@AV, which confirms its role in capturing event-level consistency. When both modules are integrated, E-CMA achieves the best overall performance across all metrics, surpassing the strong baseline CoLeaF and demonstrating the complementary benefits of CMA and EMA.

Table 3. Ablation study for E-CMA. *w/o* CMA and *w/o* EMA mean without CMA and EMA, respectively.

	Method	A	V	AV	Type@AV	Event@AV
Segment level	CoLeaF†	64.2	67.4	59.9	63.8	63.3
	w/o CMA	65.4	68.2	60.4	64.7	64.4
	w/o EMA	65.9	68.8	61.0	65.2	64.8
	E-CMA	66.1	69.9	61.7	65.9	65.4
	Method	A	V	AV	Type@AV	Event@AV
Event level	CoLeaF†	53.2	64.1	52.4	56.6	52.7
	w/o CMA	54.4	64.7	52.3	57.2	53.8
	w/o EMA	54.5	65.5	52.9	57.7	54.0
	E-CMA	54.5	66.6	53.5	58.2	54.3

4 Conclusion

In this paper, we have presented E-CMA, a novel framework for audio-visual video parsing that incorporates the cross-modal alignment and exponential moving average modules. EMA module establishes a teacher–student scheme, where the EMA teacher generates reliable segment-level pseudo labels to guide the student. CMA enforces class-aware cross-modal consistency at the segment level, enhancing audio–visual alignment. Experiments on LLP and UnAV-100 show the effectiveness of our framework. Our approach still relies on fixed strategies for pseudo label generation, which may not fully adapt to varying event distributions. In future work, we aim to develop more adaptive teacher–student update and selection mechanisms to address this issue.

5 Acknowledgements

This work was partly supported by a research scholarship from the China Scholarship Council (CSC). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

[1] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *European Conference on Computer Vision*. Springer, 2020, pp. 436–454.

[2] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Transactions on Multimedia*, vol. 25, pp. 418–429, 2021.

[3] G. Li, H. Du, and D. Hu, "Boosting audio visual question answering via key semantic-aware cues," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5997–6005.

[4] R. Guo, X. Ying, Y. Chen, D. Niu, G. Li, L. Qu, Y. Qi, J. Zhou, B. Xing, W. Yue *et al.*, "Audio-visual instance segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 550–13 560.

[5] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1326–1335.

[6] Y. Chen, R. Guo, X. Liu, P. Wu, G. Li, Z. Li, and W. Wang, "Cm-pie: Cross-modal perception for interactive-enhanced audio-visual video parsing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8421–8425.

[7] X. Jiang, X. Xu, Z. Chen, J. Zhang, J. Song, F. Shen, H. Lu, and H. T. Shen, "Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 719–727.

[8] F. Sardari, A. Mustafa, P. J. Jackson, and A. Hilton, "Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–17.

[9] J. Gao, M. Chen, and C. Xu, "Learning probabilistic presence-absence evidence for weakly-supervised audio-visual event perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[10] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 431–448.

[11] J. Gao, M. Chen, and C. Xu, "Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 827–18 836.

[12] 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 *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[13] 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 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Y.-H. Lai, Y.-C. Chen, and F. Wang, "Modality-independent teachers meet weakly-supervised audio-visual event parser," *Advances in Neural Information Processing Systems*, vol. 36, pp. 73 633–73 651, 2023.

[15] K. K. Rachavarapu, K. Ramakrishnan *et al.*, "Weakly-supervised audio-visual video parsing with prototype-based pseudo-labeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 952–18 962.

[16] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 722–34 733, 2022.

[17] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] T. Geng, T. Wang, J. Duan, R. Cong, and F. Zheng, "Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 942–22 951.

[19] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 402–419.

[20] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[21] Y. Gao, X. Sun, G. Lv, D. Yu, and S. Niu, "Reinforced label denoising for weakly-supervised audio-visual video parsing," in *International Conference on Computational Visual Media*. Springer, 2025, pp. 107–124.