MIXED IN TIME AND MODALITY: CURSE OR BLESSING? CROSS-INSTANCE DATA AUGMENTATION FOR WEAKLY SUPERVISED MULTIMODAL TEMPORAL FUSION

Yonggang Zhu[⋆‡a], *Chao Tian*[†], *Zhuqing Jiang*[⋆b], *Aidong Men*[⋆], *Haiying Wang*[⋆], *Qingchao Chen*[‡b]

[⋆]Beijing University of Posts and Communications
[†]China Xi'an Satellite Control Center
[‡]Peking University

ABSTRACT

In multimodal video event localization, we usually leverage feature fusion across different axes, such as the modality and temporal axes, for better context. To reduce the costs of detailed annotations, recent solutions explore weakly supervised settings. However, we observe that when feature fusion meets weakly supervised localization, problems can occur. It may cause "feature cross-interference", which produces a smearing effect on the localization result and can't be effectively supervised with conventional multiple instance learning loss. We verify it quantitatively on the audio-visual video parsing (AVVP) task, and propose a cross-instance data-augmentation framework, which can preserve the benefits of feature fusion while providing explicit feedbacks for feature cross-interference. We show that our method can enhance performance of existing models on two weakly supervised audio-visual localization tasks, i.e. AVVP and AVE.

Index Terms— feature cross-interference, data augmentation, weakly supervised, audio-visual localization

1. INTRODUCTION

In the era of big data, datasets with large-scale corpora of video and their audio track [1,2] enabled remarkable progress of multimodal and cross-modal video understanding applications. Research on human perception [3] and prior work on audio-visual learning [4–7] have proved the effectiveness of audio-visual integration. Also, many works exploit the correspondence between the audio and visual modalities. Examples are visual localization of sound sources [8], vision-assisted audio inpainting [9], audio-visual correspondence prediction [10, 11] and audio-visual synchronization [12], with some of them also used for self-supervised pretraining. Localization of video events from audio and visual inputs has attracted more attention from the research community as well.

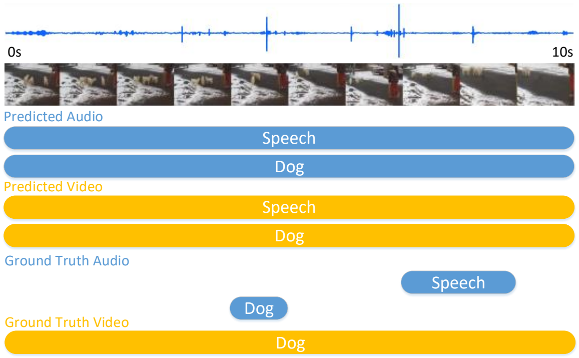

Fig. 1. An output of the model [18]. It suffers from feature cross-interference, producing a smearing effect on the output.

The AVE dataset [4] brought much interest to this field, with follow-up research such as [5, 7, 13–17]. Another dataset is the LLP dataset [18] for audio-visual video parsing (AVVP). Both datasets have temporal localization annotations.

Three challenges exist within the audio-visual event localization task. The first is to better leverage the temporal and modal context. For example, it's difficult to distinguish lawn mowers and helicopters solely from the audio modality. For this, existing works explore how to fuse the audio and visual modalities across different time steps, such as [4, 7, 14, 15].

The second challenge is to handle temporal and modality inconsistency in real-life videos, such as videos with background music or out-of-screen sound sources. For this, the AVE dataset only localizes events that are both audible and visible at the same time, and the AVVP task goes further to require separate localization labels for the audio and visual modalities. Also, [13] better models the correlations in videos with attention mechanism, and [16] tries to filter out inconsistent audio-visual pairs before fusion.

The last challenge is to reduce annotation efforts for fine-grained labeling of audio and visual streams. Tasks like AVE and AVVP adopt weakly supervised (WS) learning with "bag-level" annotations (also known as multiple instance learning, MIL) as one of their settings.

[a]Works partially done at Peking University. Supported by China NSFC (62002026), National Key R&D Program (2018YFB0505200), Beijing Advanced Discipline Construction Project (BMU2019GJJXK001) and PKU-OPPO Innovation Fund (BO202103). [b] Corresponding author.

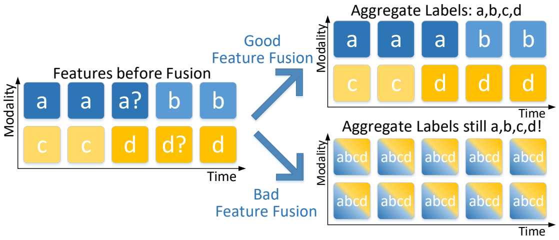

Fig. 2. An illustration of the feature cross-interference problem. "?" means this feature is not very explicit, and can be made clear by looking at its context. In weakly supervised settings, the aggregate labels won't differ whether feature fusion gets output features mixed up. Thus, the fusion step may not be motivated enough to clearly segregate output features.

However, we notice that under such weakly supervised settings, an inherent problem exists after feature fusion, which has connection with weak audio-visual and temporal consistency. We term it "feature cross-interference", which is set forth in detail in the next section and verified quantitatively through experiments on WS-AVVP.

To reap the benefits of feature fusion and overcome the pitfall of feature cross-interference, we propose a simple yet effective data augmentation framework, which can help to provide explicit feedbacks in the loss function for feature cross-interference. We evaluate our augmentation scheme on two main tasks for weakly supervised multimodal video event localization: WS-AVE and WS-AVVP. The results show that our scheme can work on many publicly available models and boost their performances. In addition, we quantitatively verify on WS-AVVP that the feature cross-interference problem has been alleviated using our method.

2. FINDINGS

We detail the concept and cause of feature cross-interference in this section. To illustrate, we take an output of the AVVP baseline model [18] as an example. As shown in figure 1, only a few seconds' audio data contain the events "speech" and "dog", yet the model predicts the two labels for all the temporal segments in the audio data. We name this problem "temporal interference". Also, the visual modality data only has dogs, yet the model hallucinates the "speech" label, which actually comes from the audio data. Similarly, we name this "modality interference". Both the modality and temporal interferences are special cases of feature cross-interference.

There may be some causes of this problem, such as unbalanced datasets or confounding factors like backgrounds. But we are particularly interested in the following reason — fine-grained localization after feature fusion with "bag-level" weak labels, which exists in both WS-AVE and WS-AVVP.

We first summarize the prevailing pipeline for weakly supervised audio-visual localization.

First, a bunch of local features is extracted from the raw video. In the task of WS-AVVP and WS-AVE, contents at each temporal segment t in each modality m ($m = 0$ for audio and $m = 1$ for visual), which we note as $\boldsymbol{x}[m,t]$, are extracted into local features $\boldsymbol{f}[m,t]$.

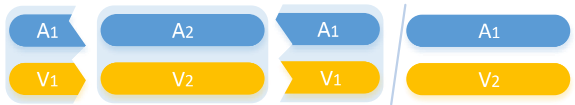

Fig. 3. Our cross-instance data augmentation on the temporal axis (left) and modality axis (right). A_1 and V_1 means the audio and visual modality of video 1. A_2 and V_2 are similar.

At the feature fusion stage, each local feature interacts and merges with each other, and outputs the context-aware feature $\hat{\boldsymbol{f}}[m,t]$ for WS-AVVP or $\hat{\boldsymbol{f}}[t]$ for WS-AVE. Each context feature $\hat{\boldsymbol{f}}[m,t]$ (or $\hat{\boldsymbol{f}}[t]$) is then individually transformed into the event probabilities $\boldsymbol{p}[m,t] \in \mathbb{R}^Y$ (or $\boldsymbol{p}[t] \in \mathbb{R}^Y$), where Y is the total number of categories. In this regard, a multi-label classification is done for each segment.

Under weakly supervised conditions, these fine-grained outputs $\boldsymbol{p}[m,t]$ (or $\boldsymbol{p}[t]$) are further pooled to predict the video-level output $\bar{\boldsymbol{p}} \in \mathbb{R}^Y$. We usually use the MIL loss [4, 5, 13, 16, 19], and the whole network is learned by checking $\bar{\boldsymbol{p}}$ against the video-level ground-truth $\boldsymbol{y}$ (union of all event labels from each modality and time step):

$$\begin{aligned} \bar{\boldsymbol{p}} &= \mathrm{Agg}(\boldsymbol{p}) \\ \mathcal{L} &= l(\bar{\boldsymbol{p}}, \boldsymbol{y}) \end{aligned} \tag{1}$$

"Agg" is the pooling operation, l is a loss criterion such as the mean square error loss or cross entropy loss.

Following the previous process, models won't be constrained to maintain disentangled context features. First, through feature fusion, the context-aware features $\hat{\boldsymbol{f}}[m,t]$ may contain irrelevant contents from other time steps or other modalities not present at modality m and time step t. Because of this, the model is highly possible to predict high probabilities in $\boldsymbol{p}[m,t]$ for these events, which are not in the segment feature $\boldsymbol{f}[m,t]$ yet within the video, causing a blurry output.

Second, this problem can't be effectively curbed through the MIL loss. As shown in figure 2, it's hard to tell from the aggregate labels whether the fine-grained labels are clear-cut (see good feature fusion in the figure) or blurred (see bad feature fusion in the figure). In MIL loss, the supervision is only applied to the aggregate predictions $\bar{p}$, making it hard to differentiate whether such feature cross-interference happens or not. Videos with weak audio-visual and temporal consistency are more prone to it. Therefore, we hypothesize that designing feature fusion components with only video-level supervision might be insufficient to tackle feature cross-interference.

3. APPROACH

We approach this problem with a cross-instance data augmentation training strategy (see figure 3) that can synthesize annotated samples with a finer granularity and help reduce feature cross-interference. Through this augmentation, the interference problem can be made visible to the loss function and explicitly optimized with modified MIL supervision.

Let the i-th video in training set $\boldsymbol{x}^i$, and its corresponding set of (video-level) ground-truth labels $\boldsymbol{y}^i$. We use $\boldsymbol{x}^i[m,t]$ to denote contents of $\boldsymbol{x}^i$ at modality m and time t.

We first select an axis for concatenation. The selected axis can either be "numerical" or "categorial". For the numerical axis, such as the temporal axis, the newly assembled video $\tilde{\boldsymbol{x}}$ and its corresponding label set $\tilde{\boldsymbol{y}}$ is formulated as:

$$\tilde{\boldsymbol{x}} = \mathrm{Cat}((\boldsymbol{x}^{i_0}[:,:t'], \boldsymbol{x}^{i_1}[:,:], \boldsymbol{x}^{i_0}[:,t':]), \mathrm{axis}=1) \tag{2}$$

$$\begin{aligned} \tilde{\boldsymbol{y}}_{t<t' \text{ or } t\geq t'+T} &= \boldsymbol{y}^{i_0} \\ \tilde{\boldsymbol{y}}_{t'\leq t<t'+T} &= \boldsymbol{y}^{i_1} \end{aligned} \tag{3}$$

where "Cat" means the concatenation operation, ":" is a python-like slicing operation, i_0 and i_1 denote two randomly selected videos in the training set with *disjoint* event labels, T is the total length of video i_0, and t' is a randomly selected split position. It is supervised with a modified MIL loss:

$$\begin{aligned} \tilde{\boldsymbol{p}}_0 &= \mathrm{Cat}((\tilde{\boldsymbol{p}}[:,:t'], \tilde{\boldsymbol{p}}[:,t'+T:]), \mathrm{axis}=1) \\ \mathcal{L} &= l(\mathrm{Agg}(\tilde{\boldsymbol{p}}_0), \tilde{\boldsymbol{y}}_{t<t' \text{ or } t\geq t'+T}) \\ &\quad + l(\mathrm{Agg}(\tilde{\boldsymbol{p}}[:,t':t'+T]), \tilde{\boldsymbol{y}}_{t'\leq t<t'+T}) \end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{p}}$ denotes the fine-grained event probabilities of the new video $\tilde{\boldsymbol{x}}$ outputted by the model. See section 2 for details. For the categorical axis, such as the modality axis, we use:

$$\tilde{\boldsymbol{x}} = \mathrm{Cat}((\boldsymbol{x}^{i_0}[0,:], \boldsymbol{x}^{i_1}[1,:]), \mathrm{axis}=0) \tag{5}$$

$$\begin{aligned} \tilde{\boldsymbol{y}}_{m=0} &= \boldsymbol{y}^{i_0} \\ \tilde{\boldsymbol{y}}_{m=1} &= \boldsymbol{y}^{i_1} \end{aligned} \tag{6}$$

with the loss function:

$$\begin{aligned} \mathcal{L} = & \, l(\mathrm{Agg}(\tilde{\boldsymbol{p}}[0,:]), \tilde{\boldsymbol{y}}_{m=0}) \\ & + l(\mathrm{Agg}(\tilde{\boldsymbol{p}}[1,:]), \tilde{\boldsymbol{y}}_{m=1}) \end{aligned} \tag{7}$$

In this way, we feed the new video $\tilde{\boldsymbol{x}}$ to the model, expecting that the model shouldn't fuse contents of $\boldsymbol{x}^{i_0}$ with those in $\boldsymbol{x}^{i_1}$ since they aren't correlated and share no common labels.

If the model suffers feature cross-interference and erroneously mix contents in $\boldsymbol{x}^{i_0}$ with those in $\boldsymbol{x}^{i_1}$, it is highly possible that context-aware features of $\boldsymbol{x}^{i_0}$ may be mis-classified to contain events in $\boldsymbol{y}^{i_1}$, while the features of $\boldsymbol{x}^{i_1}$ being misclassified to contain events in $\boldsymbol{y}^{i_0}$.

Thus, if we pool indexes corresponding to $\boldsymbol{x}^{i_0}$ and $\boldsymbol{x}^{i_1}$ separately, and supervise with their respective video-level labels, such mixing will be discouraged.

In addition, it can be observed that the annotations have a finer granularity after this augmentation scheme. For example, after temporal augmentation, the newly formed data has separate labels for $t < t' \cup t \geq t' + T$ and $t' \leq t < t' + T$. We note that under modality augmentation, the synthesized label may be noisy, since some events may only occur in one modality for some videos. To handle it, we follow the steps in [18] and apply label smoothing. Our proposed training strategy is "plug-and-play", which means it is compatible with many weakly supervised audio-visual localization models. Note that our algorithm has connection with data augmentation for Monte Carlo parameter estimation [20] with the latent variables $\boldsymbol{z}$ the concatenation parameters. It's possible to integrate our method with other strategies, such as resampling or deconfounding. Also, our method does not introduce any extra computational costs at inference time or any extra trainable parameters.

4. EXPERIMENT AND ANALYSIS

4.1. Dataset used

The Audio-Visual Event (AVE) dataset [4] and LLP dataset (for WS-AVVP) [18] are used in our evaluation, which are both popular datasets for audio-visual event localization. The AVE dataset consists of 4143 videos and 28 categories, while the LLP dataset comprises 11849 videos and 25 categories. Each video in both datasets is 10s long. The AVE dataset annotates audio-visual events for each 1s segment on the temporal axis. The LLP dataset only has video-level labels in the training set. However, in the test and validation set, the event labels for each 1s segment are available, as well as in which modality of the segment each event comes from.

4.2. Performance on WS-AVVP and WS-AVE

We evaluate the performance of the proposed cross-instance augmentation scheme on WS-AVVP and WS-AVE.

On WS-AVVP, we adopt the base model as in [18]. It features a transformer-based [21] feature fusion module (namely the HAN layer). We train it with an alternation strategy where the original dataset is trained for 2 epochs, followed by 1 epoch of augmentation on the modality axis and another epoch on the temporal axis. We use the Adam optimizer with a base learning rate of 3e-5. The overall framework is trained for a total of 60 epochs and we save the best performing models based on the validation set. We use the same audio and visual features as [18] to ensure fairness. The metrics are based on F-score, detailed in [18]. Components in [18] such as individually guided loss and label smoothing are retained.

On WS-AVE, we apply temporal augmentation on two models: the AVEL model [4] and the PSP model [16]. The audio and visual features come from [4]. We use the Adam optimizer with a base learning rate of 1e-3 and train for 300 epochs. The metric is localization accuracy as in [4].

The results on WS-AVVP are shown in table 4.2. We can see our method outperforms the original model on *all* metrics. Furthermore, we perform an ablation on our augmentation scheme based on augmenting either the modality axis or the temporal axis. Our method still enhances the base model when settings of the HAN are changed to 2L2H (L: layers, H: heads). The average improvement on Segment Ty is 2.2±0.4

Metric	Segment					Event				
	A	V	A-V	Ty	Ev	A	V	A-V	Ty	Ev
Original	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
+M	60.2	55.1	49.8	55.0	**57.1**	**51.8**	**51.1**	43.0	48.7	**48.8**
+T	60.4	54.4	50.5	55.1	55.7	51.3	48.6	43.1	47.7	47.8
+M+T	**60.7**	**55.2**	**51.8**	**55.9**	56.6	**51.8**	50.3	**44.4**	**48.8**	48.2
2L2H	57.3	52.8	48.3	52.8	53.6	48.1	48.9	42.3	46.5	45.9
+M+T	**61.6**	**54.0**	**50.8**	**55.5**	**58.2**	**52.6**	**49.0**	**44.4**	**48.7**	**49.1**

Table 1. Results on WS-AVVP (+M+T) with ablations on augmentation for the modality axis (+M) and temporal axis (+T). A: Audio; V: Visual; Ty: Type@AV; Ev: Event@AV.

for all combinations of L from 1∼4 and H from {1,2,4}. On WS-AVE, results in table 4.2 show that our method can boost the performance of both the AVEL and PSP model.

Method	Accuracy	Method	Accuracy
AVEL [4]	66.7	AVIN [7]	69.4
AVEL w/ aug	68.4	AVT [22]	70.2
AVSDN [5]	67.3	CMRA [15]	72.9
CMAN [13]	70.4 [16]	PSP [16]	72.7*
AVRB [14]	68.9	PSP w/ aug	**73.2**

Table 2. Performance comparison on WS-AVE [4]. * denotes our reproduced results. "w/ aug" means "with augmentation".

4.3. Existence and reduction of feature cross-interference

To quantitatively examine the existence of feature cross-interference, we first conduct experiments on how the models perform on visual and audio paired data with mismatched semantics. We extract two subsets from the test set of the LLP dataset, one with *non-identical* labels for the audio and visual modalities (623 instances), while the other with *non-common* audio and visual labels (161 instances). These subsets are evaluated on the original WS-AVVP model [18]. The results are listed in table 4.3. It can be seen that the model performance undergoes significant degradation for videos with different cross-modal semantics. We hypothesize that under the mismatched semantic circumstances, the feature cross-interference problem is more prone to happen. However, it can be observed that after applying our augmentation scheme, the interference problem has been alleviated significantly.

Next, we adopt four additional metrics to evaluate the feature cross-interference problem across different modalities. These metrics are Audio False Positive from Visual Output (A_FP_VO), Visual False Positive from Audio Output (V_FP_AO), Audio unique label Leakage to Visual output (A_Leak_V), and Visual unique label Leakage to Audio output (V_Leak_A). They can be formulated as follows:

$$\text{A\_FP\_VO} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{sum}[(o_a^i * (1-g_a^i) * o_v^i]}{\text{sum}[o_a^i * (1-g_a^i)]} \tag{8}$$

$$\text{V\_FP\_AO} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{sum}[(o_v^i * (1-g_v^i) * o_a^i]}{\text{sum}[o_v^i * (1-g_v^i)]} \tag{9}$$

$$\text{A\_Leak\_V} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{sum}[g_a^i * (1-g_v^i) * o_a^i * o_v^i]}{\text{sum}[g_a^i * (1-g_v^i) * o_a^i]} \tag{10}$$

$$\text{V\_Leak\_A} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{sum}[(g_v^i * (1-g_a^i) * o_v^i * o_a^i]}{\text{sum}[g_v^i * (1-g_a^i) * o_v^i]} \tag{11}$$

where N is the total number of evaluation instances, o_a^i is the audio modality output of the i-th instance, o_v^i is the visual modality output of the i-th instance, g_a^i is the audio ground-truth for the i-th instance, g_v^i is the visual ground-truth for the i-th instance. They all have the same shape $\{0,1\}^{Y \times T}$, with Y the number of classes and T the temporal length of the video. "*" denotes element-wise multiplication. The results of these metrics with the original AVVP model [18] on the test set are in table 4.3. Also, we can see the severity of modality interference is alleviated after augmentation.

Metric	Non-subsetted	Non-identical	Non-common
Segment A	60.7 (60.1)	53.3 (52.0)	47.5 (45.6)
Segment V	55.2 (52.9)	44.8 (42.2)	34.5 (29.2)
Segment A-V	51.8 (48.9)	41.7 (37.7)	34.3 (25.4)
Segment Ty	55.9 (54.0)	46.6 (44.0)	38.8 (33.4)
Segment Ev	56.6 (55.4)	46.5 (44.3)	41.8 (37.0)
Event A	51.8 (51.3)	43.6 (40.7)	39.9 (37.0)
Event V	50.3 (48.9)	40.3 (38.3)	33.0 (28.4)
Event A-V	44.4 (43.0)	35.2 (31.9)	34.3 (25.4)
Event Ty	48.8 (47.7)	39.7 (37.0)	35.8 (30.3)
Event Ev	48.2 (48.0)	38.0 (35.9)	32.7 (30.1)

Table 3. Performance on videos with weak audio-visual correspondence before and after augmentation. Original results are shown in parentheses. Abbreviations are as table 4.2.

Metric	Result	Metric	Result
A_FP_VO	68.7%(79.2%)	V_Leak_A	70.9%(91.7%)
V_FP_AO	60.1%(85.9%)	A_Leak_V	52.6%(80.5%)

Table 4. Results of the interference metrics before and after augmentation. Original values are shown in parentheses.

5. CONCLUSION

We discover the problem of feature cross-interference which could happen in feature fusion under many weakly supervised multimodal video event localization tasks, and propose a cross-instance data augmentation scheme to alleviate the problem. We verify the existence of this problem quantitatively on the task of WS-AVVP and show that our augmentation can significantly alleviate it. Finally, we test the effectiveness of the augmentation scheme on WS-AVVP and WS-AVE, which shows our method can be applied to and push further the performance of many existing models.

6. REFERENCES

[1] J. F. Gemmeke, D. P. W. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[2] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[3] J. K. Bizley, R. K. Maddox, and A. K. Lee, "Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms," *Trends in Neurosciences*, vol. 39, no. 2, pp. 74–85, feb 2016.

[4] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[5] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[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.

[7] J. Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple sound sources localization from coarse to fine," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[9] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[12] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[13] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[14] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3893–3901.

[16] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[17] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[19] Y. Wang, J. Li, and F. Metze, "A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[20] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of monte carlo methods for parameter estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, pp. 1–62, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[22] Y.-B. Lin and Y.-C. F. Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.