ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747579

NAVIGATING AUDIO-VISUAL EVENT DETECTION ACROSS MISMATCHED MODALITIES

Guangwei Li, Xuenan Xu, Mengyue Wu†, Kai Yu†

X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence
AI Institute, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Previous audio-visual (AV) alignment mainly focuses on frame-level synchronization while neglecting clip-wise matching. We focus on AV parsing on fully unconstrained data where the audio and visual events do not necessarily co-present. A video-enhanced Audioset dataset is provided to investigate parsing on such a mismatching setting, with 376 events included. To our knowledge, this is the first time where AV event parsing and detection are inspected on a clip-wise matching scenario. Experiments show that our proposed method largely improves video parsing accuracy on tagging and detection. Further, a parsing model pre-trained on our dataset can assist in accurately locating audio-visual syncing time spans.

***Index Terms*—** audio-visual event detection, clip-level mismatch, weakly-supervised, multimodal

1. INTRODUCTION

Auditory and visual cues are commonly complementary to each other on the condition that they co-exist during a time span. On one aspect, audio can assist video modality in traditional computer vision tasks, including action recognition [1] and video question answering [2]. Audio information provides crucial information to identify action or scenes, thus leading to better system performance. On the other aspect, visual information (both image and video) prove helpful to traditional speech and audio processing tasks like audio tagging [3, 4, 5], source separation [6, 7] and speaker verification [8, 9]. Introducing visual information enables the system to better recognize sound events or speech.

Apart from the works mentioned above, where one modality is often used to assist another modality, efforts have been made to dive into the connections and differences between both modalities [10, 11, 12]. However, visual and auditory description systems are different by nature as they belong to different senses, which leads to two levels of mismatch of audio-visual (AV) events, *clip-level co-presence* and *frame-level co-occurrence*. A clip-level co-present event appears both in the audio and visual modality in one single clip, while a frame-level synchronized event happens simultaneously in both modalities. As shown in Figure 1, "Speech" only appears in audio modality, while "Vehicle" appears in both modalities. Here, AV mismatch of "vehicle" is at frame-level as the time spans are different; however, "speech" is a clip-level mismatch AV event.

Most previous AV multimodal works concentrate on the latter condition, where the frame-level AV interaction is investigated. Audio-visual event detection (AVE) [10] focuses on frame-level synchronized AV events and proposes a novel dual multimodal residual network to address AV localization. Audio-visual parsing (AVP) [11] task is proposed to disentangle frame-level asynchronous scenarios by separately detecting audio-only, video-only, and audio-visual events. Taking the frame mismatch into consideration, the AVP dataset selects events that are present in both modalities, which is limited to as few as 25 pre-defined events. However, AV mismatch not only exists on time span but also at clip-level. Generally, frame-level synchronization indicates a clip-level one. Nevertheless, the clip-level mismatch can lead to further confusion, which broadly occurs across a large variety of events in real life. For instance, many documentaries have voice-over in the audio channel with no humans present in the image. In this way, such label mismatch phenomenon greatly limits audio-visual research but has rarely been investigated.

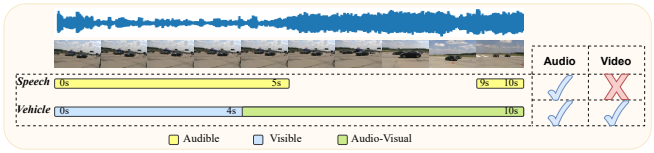

Fig. 1. An example of event modality clip-level-mismatch. Vehicle can be seen and heard in the clip while we cannot see but only hear people speaking and cheering in the background. Here, *vehicle* is a frame-level AV mismatched event while *speech* is clip-level mismatched, which cannot be processed by previous parsing network.

We innovatively take clip-level AV mismatch into consideration and conduct AV event parsing and detection on fully unconstrained data. As many as 376 AV events can be detected, largely outnumbered previous AV datasets. A Video-Enhanced Balanced Audioset (VEBA) dataset is provided, with clip-wise video presence labels added to the existing weakly-labeled sound event dataset. Compared to previous audio-visual tasks, our study enables audio-only, video-only, and audio-visual event (AVE, fully synchronized event) detection in real-world scenarios. Experiments show that with extra video labels, higher precision is achieved in video event and AVE detection on both tagging and detection. Analysis on the out-domain dataset AVE suggests that our system can automatically screen and detect synchronized audio-visual events, which may further assist research on other audio-visual tasks that require simultaneity.

2. VEBA: VIDEO ENHANCED BALANCED AUDIOSET DATASET

In this section, we provide details towards VEBA selection, labeling procedures as well as comparisons with previous datasets. We

† Mengyue Wu and Kai Yu are the corresponding authors.

This work has been supported by National Natural Science Foundation of China (No.61901265), State Key Laboratory of Media Convergence Production Technology and Systems Project (No.SKLMCPTS2020003) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

choose audio event datset as our startpoint since there are profoundly more audio events defined (i.e. 527 in Audioset [13]) than visual events/objects (i.e. 101 in UCF-101 [14]). A merging and selection strategy is adopted to ensure that each event has ample samples in both modalities. For example, many event sub-categories in Audioset are quite similar and require expertise knowledge [15]. We merge the sub-categories into their parent node. Finally, 376 events are obtained, which is a subset of the original 527 labels. We obtain the videos (accompanied with the corresponding audios) from balanced Audioset (originally encompassing) and provide clip-level weak annotations for the video modality. Annotators are invited to label the video presence of each audio event, namely whether this event occurs in video or not. We expect this supplementary in ground truth label will help the system better learn the alignment as well as the mismatch between modalities. The original weak labels of Audioset clips demonstrate the existence of the labeled events in audio modality or, in other words, audible. However, whether this event is visible is unknown in these Youtube-originated wild videos; hence the current video presence labeling is important. For instance, for an video clip with original label "Speech, Music and Vehicle". We ascertain that all three events are audible, according to the original Audioset labels. With True/False labels provided for the three events' presence in video modality, we acquire the weakly-labeled AV status for each event. At last, VEBA includes 18,765 video clips with corresponding audios. A total of 376 event labels is included, with 36,203 audio events and 17,742 video events obtained. 2k videos from Audioset evaluation set are selected as our test set, where we separately annotate the onset and offset of events in audio and visual modality with a time resolution of one second. We split about 10 percent of the clips (1600 precisely) as the VEBA validation set, so the number of clips in the training set is 15,278. The final data distribution of VEBA dataset can be seen in Table 1.

Table 1. A summary of data in VEBA dataset, including detailed clip distribution of training, validation, test set and label count (Audio and Visual).

Split	# Clips	# Weak labels		#On-offset pairs	
		A	V	A	V
Train	15278	30422	14902	×	×
Val	1600	3110	792	×	×
Test	1887	2671	2048	3166	2282
Total	18765	36203	17742	3166	2282

A brief summary of the weak audio and video labels in VEBA is illustrated in Figure 2. The horizontal axis is the number of the audio labels of an event category, while the vertical axis is the number of video labels of an event category. The horizontal-vertical position in the figure shows the frequency of an event in the VEBA dataset. For example, "Music" and "Speech" appear the most, while "Ringtone" and "Echo" appear the least. The color of the event indicates the ratio of video label number and audio label number of an event. Blue events are more inclined to been "seen" while red events are usually "heard". The top three visible events are "Vacuum Cleaner", "Blender" and "Spray", while the top three audible events are "Plop", "Jingle tinkle" and "Echo", which is consistent with common knowledge.

Table 2 shows the comparison between our VEBA dataset with previous audio-visual event datasets such as LLP dataset [11] and AVE dataset [10]. As mentioned above, these datasets exhibit certain constraints, focusing on events that happen simultaneously in audio and visual modalities. The event number is 28 in AVE and 25 in LLP,

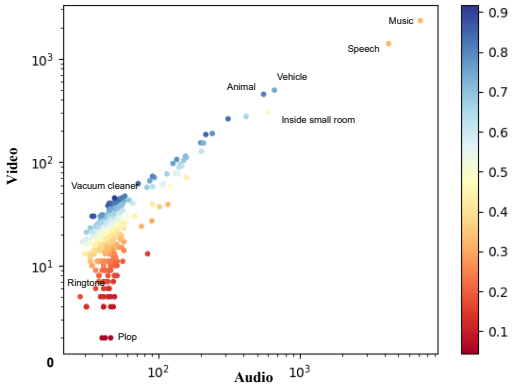

Fig. 2. A summary of our weak labels in Audioset balanced dataset. Each dot represents an event from our VEBA dataset. The horizontal axis represents the number of event appears in audio labels while the vertical axis represents the number in video labels. The color of an event represents the ration of video and audio.

which is far from the number of audio-visual events in real life. In order to better facilitate real-life audio-visual event parsing research, we provide as many as 376 event categories with both frame-level and clip-level mismatched situations included.

Table 2. A comparison between our VEBA dataset with the previous audio-visual dataset. AV-M means whether the dataset include the mismatched audio and video events in a clip.

Dataset	Clips	Events included	AV-M
AVE Dataset [10]	4143	28	×
VGG-Sound [16]	**200k**	309	×
LLP Dataset [11]	11849	25	✓
VEBA Dataset	18765	**376**	✓

3. AUDIO-VISUAL PARSING MODEL

In order to predict the onset and offset of the events in different modalities, we use a parsing model shown in Figure 3 to learn the modality alignment from audio and video. The whole framework is explained as follows: 1) Single single modal feature extractors and encoders; 2) A hybrid attention network combining the information and leveraging multimodal contexts; 3) Attentive pooling along with training losses.

Single modal extractor and encoder To better capture modality characteristics, we use pre-trained networks to extract the features from audio and video. As for the audio, we use a convolutional recurrent neural network (CRNN) named L-CDur, based on CDur [17]. It is trained on the unbalanced ($\approx 5000h$) subset of Audioset. We remove the last layer of the model to extract the audio feature. Regarding the video, we use the combined feature from Efficientnet-b6 [18, 19] and 3D Resnet [20], pre-trained on Imagenet [21] and Kinetics [22]. We transform the two video features into the same dimension (512) and concatenate them. A fully connected (FC) layer is added to project the features from different modalities into the same size. We further use a multi-head self-attention-based [23] encoder to learn the sequential information in

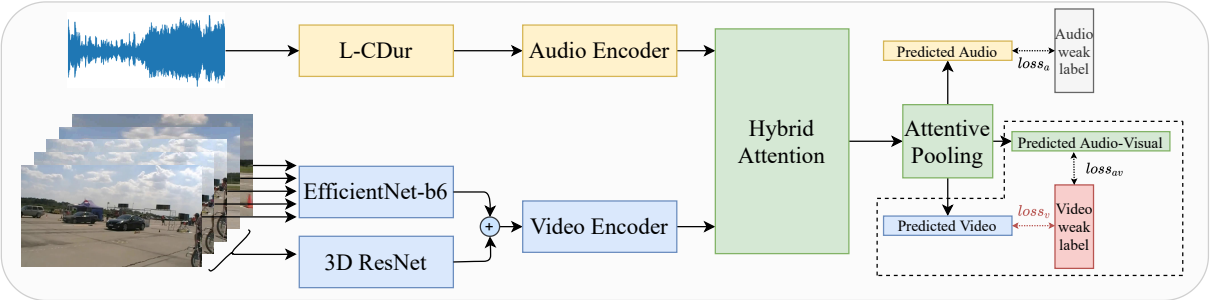

Fig. 3. The whole structure of our audio-visual network. For an input clip, we use pre-trained single modal extractors and encoders to separately extract the audio and video feature. Then a hybrid attention module combines the information and leverages multimodal contexts. An attentive pooling is used to predict individual outputs for events in audio, video and audio-visual modalities. The loss is to calculate the distance between the weak ground truth labels and the predicted labels.

both modalities better.

Hybrid attention network After obtaining the independent information from audio and video, we endeavor to combine the information from both modalities. We use a hybrid attention network [11] with both self-attention and cross-modality attention. This hybrid attention network (HAN) can adaptively learn which modality to attend for each audio or visual snippet.

Attentive pooling The output of HAN is the temporal aggregated embeddings $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^T$ of audio and video. After a shared FC layer and a sigmoid function, the audio and visual probabilities for each event are obtained.

$$p_a^t = sigmoid(FC(\hat{f}_a^t)) \tag{1}$$

$$p_v^t = sigmoid(FC(\hat{f}_v^t)) \tag{2}$$

Audio and video clip level probabilities $\boldsymbol{p_a}$ and $\boldsymbol{p_v}$ are estimated by summing the frame-level probabilities $p_{a/v}^t$ along the time axis. To predict the clip-level audio-visual event probability $\bar{\boldsymbol{p}}$, an attentive pooling [11] is used to judge which modality is more trustworthy at each moment:

$$\bar{\boldsymbol{p}} = \sum_{t=1}^{T} \sum_{m=1}^{M} (W_{tp} \odot W_{av} \odot P)[t, m, :] \tag{3}$$

where $\odot$ is the element-wise multiplication, m is the modality index and M equals 2 here for audio and video modalities. W_{tp} and W_{av} are temporal and audio-visual attention calculated from $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^T$. P is the concatenation of p_a^t and p_v^t.

Loss function As seen in Section 2, both audio and video weak labels are available for training. We denote the audio ground truth label as $\boldsymbol{y_a}$ and the video ground truth label as $\boldsymbol{y_v}$. Note that according to our labeling rule, all of the video event labels of a given clip is included in its audio event labels. Our audio-visual ground truth label is also $\boldsymbol{y_v}$. The target of our model is to optimize the following loss:

$$\mathcal{L} = \mathcal{L}_a + \mathcal{L}_v + \mathcal{L}_{av} = CE(\boldsymbol{p_a}, \boldsymbol{y_a}) + CE(\boldsymbol{p_v}, \boldsymbol{y_v}) + CE(\bar{\boldsymbol{p}}, \boldsymbol{y_v}) \tag{4}$$

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental Setup

Data preparation As mentioned in Section 2, our training set includes 16,878 video clips with corresponding audios. Our audio feature extractor takes log-mel spectrum as input, the same configuration as CDur [17].

System Configurations We sample the videos with a sample rate of 8fps, so a 10-second long video is divided into 80 frames of images. Both the audio and video features extracted (from L-CDur, Efficient-b6, and 3D Resnet) are converted to 512-D so as to keep the temporal consistency in both modalities. Adam optimizer is used for training the system with an initial learning rate of $2e-4$ and a decay of 0.1 after every 20 epochs. We choose the model with the best accuracy (mAP) in video and audio-visual tagging on the validation set as our final model.

Evaluation metrics Following previous parsing work [11], for all three modalities (audio, video, and audio-visual), we evaluate the performance of both clip-level event tagging and temporal-level event detection. For tagging, we calculate mean average precision (mAP), while for detection, we calculate segment-level and event-level metrics (F-score).

4.2. Results

Methods in comparison We provide a baseline where clip-level mismatch is neglected (leave out the structure in dotted frame in fig. 3), as previous AV research commonly assumes AV co-presence and combines the information from audio and video. Here, the model output is regarded as a unified representation for audio, video and audio-visual. Tian *et al.* [11] is the work that firstly proposed AVP to disentangle asynchronous AV events, therefore taken as another compared method. We include this comparison to show that in addition to frame-level synchronization, clip-level mismatch plays an important role in AV parsing.

Tagging performance As presented in Table 3, offering extra visual-modality event labels in training improves both video and audio-visual event tagging performance, achieving a 2.7% and 3.2% increase in mean average precision in two modalities, respectively. The audio tagging performance is unaffected. High tagging accuracy shows the effectiveness of our model in differentiating clip-level event mismatch and synchronization.

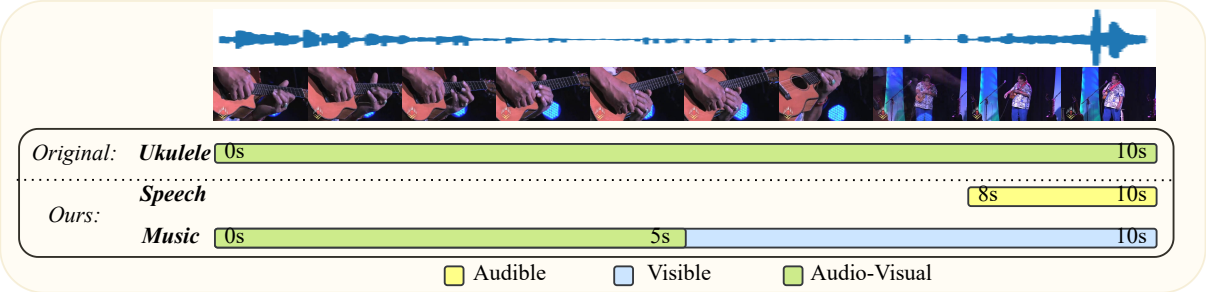

Fig. 4. An example of the visualization of our audio-visual parsing result in a video from AVE Dataset. The original event label of this clip is "Ukulele" from 0-10 seconds. However, the man in the video only plays ukulele from 0 to about 6 seconds. When the man stands up holding the musical instrument, the crowd begins to cheer and shout. Since we categorize all the children of "Music" as "Music", our model shows that "Music" happens in audio-visual from 0-5 seconds and only visible in the rest, a more precise annotation than the original AVE label.

Table 3. Tagging performance of our proposed model compared to other methods.

Configuration	Tagging (mAP(%)) ↑		
	Audio	Video	Audio-Visual
(1) Baseline	53.50	38.25	38.25
(2) Tian *et al.* [11]	54.45	41.44	43.19
(3) Ours	54.23	**43.17**	**46.37**

Event detection performance Audio, video, and audio-visual's onset and offset detection results can be seen in Table 4 . Similar to the event tagging result, our model achieves a significantly better result in event parsing tasks, especially in audio and audio-visual modalities, with the most considerable increase of 21.3% (from 18.3 to 39.6), which is a relative 116.39% improvement in event-level F-score in video event detection. High event detecting accuracy indicates that by taking clip-level mismatch into consideration, frame-level synchronization performance is greatly enhanced.

Table 4. Audio-Visual parsing (or event detection) performance compared to other methods.

Configuration	Segment-level (F-score%) ↑			Event-level (F-score%) ↑		
	A	V	A-V	A	V	A-V
(1) Baseline	38.6	31.9	42.2	33.2	31.9	42.1
(2) Tian *et al.* [11]	40.4	23.7	29.0	35.4	18.3	24.9
(3) Ours	**41.0**	**41.5**	**50.0**	**35.9**	**39.6**	**48.2**

4.3. Analysis

To further illustrate the effect of the extra weak video event labels in our dataset, we vary the proportion of the video labels used. When using only part of the video labels, we randomly choose a proportion of videos and provide them with newly-annotated video labels. For the clips not chosen, we use the original audio event labels as the video event labels, same with the original Audioset. The result is shown in Figure 5, note that we only select audio-visual F-score for demonstration. There is a steady improvement of both segment-level and event-level F-score of event parsing in video and audio-visual modalities as we vary the video labels used from 20% to 100%.

Further application Apart from evaluating our model's performance on VEBA, we also apply our model to other audio-visual datasets like AVE. As mentioned in [10], theoretically, all of the events labeled in AVE dataset happen simultaneously in audio and video. However, there still exists a frame-level mismatch between the original audio-visual event labels. The example shown in Figure 4 fairly illustrates the effectiveness of our model in audio-visual parsing and the ability to distinguish the mismatch (clip-level and frame-level) between audio and video modalities.

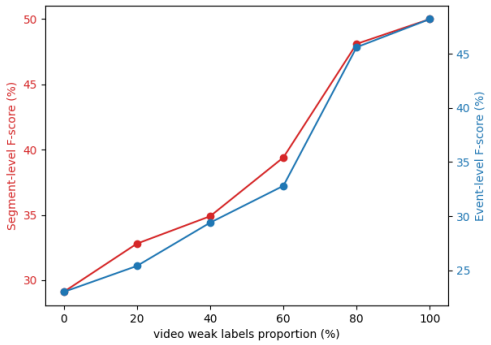

Fig. 5. Both the segment-level and event-level F-score of event parsing steady increases as we vary the proportion of weak video labels used from 20% to 100%, showing the effectiveness of our data to audio-visual parsing.

Further applications may include an exquisite revise to audio-visual event labels in the existing datasets or detecting and selecting synchronized audio-visual segments from videos in the wild for future pre-training. The demos are available online[1].

5. CONCLUSION

In this work, we further investigate the audio-visual parsing task and focus on the frame-level and clip-level mismatch of real-life events. We enrich the number of AV events included (25 → 376) and provide extra video weak event labels in our proposed VEBA dataset, on which we train and evaluate the performance of an AV parsing model. Considering clip-level mismatch not only enables AV parsing on significantly more events from entirely unconstrained data but also dramatically improves frame-level synchronization estimations. Our model achieves better performance in both event tagging (clip-level) and detection (segment-level and event-level) in video and audio-visual modalities than in previous work. Further applications may include refining audio-visual labels and selecting synchronized event segments in videos in the wild.

[1] https://ligw1998.github.io/multimodal.html

6. REFERENCES

[1] Jing Liu, Xinxin Zhu, Fei Liu, Longteng Guo, Zijia Zhao, Mingzhen Sun, Weining Wang, Hanqing Lu, Shiyu Zhou, Jiajun Zhang, and Jinqiao Wang, "OPT: Omni-Perception Pre-Trainer for Cross-Modal Understanding and Generation," 2021.

[2] Hung Le, Doyen Sahoo, Nancy F. Chen, and Steven C.H. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 5612–5623, 2020.

[3] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira, "Perceiver: General Perception with Iterative Attention," 2021.

[4] Haytham M. Fayek and Anurag Kumar, "Large scale audio-visual learning of sounds with weakly labeled data," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2021-Janua, pp. 558–565, 2020.

[5] Yuan Gong, Yu-An Chung, and James Glass, "PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation," pp. 1–15, 2021.

[6] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[7] Chenda Li and Yanmin Qian, "Listen, watch and understand at the cocktail party: Audio-visual-contextual speech separation.," in *Interspeech*, 2020, pp. 1426–1430.

[8] Yanmin Qian, Zhengyang Chen, and Shuai 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.

[9] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Multi-modality matters: A performance leap on voxceleb.," in *INTERSPEECH*, 2020, pp. 2252–2256.

[10] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263.

[11] Yapeng Tian, Dingzeyu Li, and Chenliang Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 436–454.

[12] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2002–2006.

[13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[15] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[16] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[17] Heinrich Dinkel, Mengyue Wu, and Kai Yu, "Towards Duration Robust Weakly Supervised Sound Event Detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 887–900, 2021.

[18] Hao Tan and Mohit Bansal, "LXMert: Learning cross-modality encoder representations from transformers," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 5100–5111, 2020.

[19] Mingxing Tan and Quoc V. Le, "EfficientNetV2: Smaller Models and Faster Training," 2021.

[20] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5999–6008.