Event-Specific Audio-Visual Fusion Layers: A Simple and New Perspective on Video Understanding

Arda Senocak[1*] Junsik Kim[2*] Tae-Hyun Oh[3] Dingzeyu Li[4] In So Kweon[1]
[1] KAIST [2] Harvard University [3] Dept. of EE, POSTECH [4] Adobe Research

Abstract

To understand our surrounding world, our brain is continuously inundated with multisensory information and their complex interactions coming from the outside world at any given moment. While processing this information might seem effortless for human brains, it is challenging to build a machine that can perform similar tasks since complex interactions cannot be dealt with a single type of integration but require more sophisticated approaches. In this paper, we propose a new simple method to address the multisensory integration in video understanding. Unlike previous works where a single fusion type is used, we design a multi-head model with individual event-specific layers to deal with different audio-visual relationships, enabling different ways of audio-visual fusion. Experimental results show that our event-specific layers can discover unique properties of the audio-visual relationships in the videos, e.g.*, semantically matched moments, and rhythmic events. Moreover, although our network is trained with single labels, our multi-head design can inherently output additional semantically meaningful multi-labels for a video. As an application, we demonstrate that our proposed method can expose the extent of event-characteristics of popular benchmark datasets.*

1. Introduction

Real-world events around us consist of different multisensory signals and their complex interactions with each other. In-the-wild videos of real-life events and moments capture a rich set of multi-modalities and their complex interactions therein. Thus, it is essential to leverage multisensory information for better video understanding, but their diversity and complex nature make it challenging. For instance, even though audio and vision signals are congruent, the way they relate are different. All these events have different types of characteristics (such as uni-modal types of vision only and audio only, and multi-modal types of continuous, instant, rhythmic, *etc*., as shown in Figure 1) which we call them as *event types*. That is, understanding video contents requires to properly deal with such diverse and complex associations and relationships. However, surprisingly, this has been overlooked by prior audio-visual recognition research.

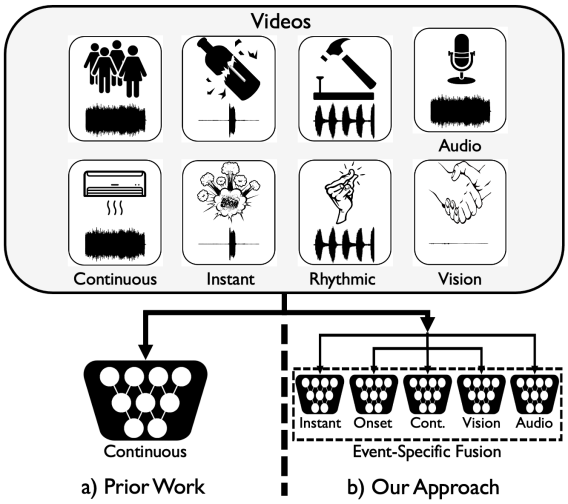

Figure 1. **A conceptual difference between prior approaches and our event-specific fusion.** Multi-modal events can be based on various forms in in-the-wild videos; while some events might have continuous temporal correspondence between visual changes and accompanied audio, the others may have rhythmic, repetitive audio-visual events or a few isolated instant moments, *e.g*., a person snapping her fingers rhythmically with background music rhythm; the air conditioner is blowing continuously; or a volcano explodes in the footage. Despite these diversities, prior approaches use a single type of one-size-fits-all fusion methods with barely considering diverse event-types. In contrast, we use multiple event-specific layers for better video understanding.

There have been vast efforts, *e.g*., [28, 40, 53, 59, 26, 31, 62], to implement a machine perception for multi-modal video analysis. A common paradigm in fusion methods [4, 65, 28, 59, 20] for audio-visual learning is to globally-pool both modalities over an entire sequence. They model multi-modal fusion mechanisms under the assumption that

* Equal contribution.

Acknowledgment. T.-H. Oh was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub, 50%; No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities, 50%).

all audio-visual events are highly correlated, aligned in time, and continuous throughout a video. This assumption is inherently injected into models once a global pooling-based fusion is used as a design choice in the architectures. This uniform assumption is flawed in its ability to model real-world multisensory events,[2] which smooths out sparse important moments and results in wrong prediction. In a race event, if the "crowd" sound is dominant instead of the "car engine" sound, the global-pooling based fusion method smooths out the "car engine" sound which may be an important signal for correct recognition. We argue that different cross-modal relationships are overlooked in these prior methods using one single type of one-size-fits-all fusion mechanism, and audio-visual video events have multi-label nature.

We present a new perspective that incorporates multiple types of audio-visual relationships to improve video understanding. Our development is motivated by the way humans perceive the world: Humans are spontaneously capable of combining relevant heterogeneous signals from the same events or objects, or distinguishing a signal from one another if the source events of the signals are different. Such multisensory integration has been widely studied in cognitive science [25, 49, 50]. Inspired by this study, we propose a simple approach consisting of multiple types of fusion layers: individual modality layers and audio-visual event-specific layers. However, it is challenging to identify and develop all possible types of integration. Thus, we postulate that most of the existing events may be effectively spanned by combinations of a few dominant event types identified by the prior cognitive studies. Each layer is designed to look for different cross-modal interactions and characteristics in videos such as audio-only, visual-only, continuous, onset, and instant event layers. This leads to our multi-head design consisting of the simple event-specific layers with a proper feature selection mechanism according to cross-modal interactions.

Our experimental results show that our method improves video classification performance and also enables reliable multi-label prediction by the multi-head design. Moreover, our proposed model leads to better interpretability of videos such as understanding audio and visual signals independently or jointly based on the characteristics of events as well as providing naïve modality confidence scores. This allows us to conduct interesting analyses of existing datasets and potential applications such as multi-labeling, category-wise and dataset-wise event characteristic analysis, and sound localization. We summarize our main contributions as follows:

- We propose a new audio-visual integration method with simple event-specific layers to enable a model to understand different characteristics of audio-visual events.
- Our analyses verify each event-specific layer captures different properties of audio-visual events that result in performance improvement for video classification.
- By virtue of the simplicity, we demonstrate the interpretability of our proposed event-specific layers that is useful in various applications: dataset event-characteristic analysis, missing label detection, and dataset retargeting.
- We will release the Multi-labeled VGGSound dataset used for our multi-label evaluation, which is a partial subset of around 1200 videos annotated by 12 subjects.

[2]Let us consider the car video in Figure 3. A group of spectators talks for a long time while recording the scene before the race car passes, but the only useful moment for audio-visual integration is the short moment the car passes by, *i.e.*, instant correlation. This event is not properly understood by capturing its global context with a uniform assumption. The other association types of multi-sensory events are illustrated in Figure 1.

2. Related Work

Audio-Visual Representation Learning. Recent years have witnessed significant progress in audio-visual learning and some used audio or visual information as a supervisory signal to the other one [7, 42, 41] or leverage both of them in self-supervised learning to learn general representations assuming that there is a natural correspondence between them [4, 5, 22, 8, 28, 40, 36, 32, 45, 44, 2]. Self-supervised learning methods use different tasks such as correspondence [4, 5, 8], synchronization [28, 40] or clustering [3]. Furthermore, some other methods use audio-visual multimodal signals as self-supervision to cluster or label the unlabelled videos [6, 3]. These existing approaches assume that multisensory data is always semantically correlated and temporally aligned. As a result, they apply simple fusion techniques such as concatenation or average pooling. However, in real-world videos, multisensory data are not always naturally co-occurring. Our work investigates more diverse multisensory relationships and proposes different integration approaches in audio-visual events. Different than other existing works, Morgado *et al*. [34] explore faulty negative and positive samples, which are semantically non-corresponding samples, in contrastive learning to obtain a higher representation quality. Our work deals with *event-level* temporal correspondence which is different from *instance level* (semantic mismatch) correspondence studied in [34].

Audio-Visual Activity Recognition. Various deep learning approaches have been proposed to improve action recognition accuracy by incorporating audio as a complementary modality [31, 29, 26, 62, 19, 59]. While most of the existing works simply concatenate audio and visual features, distillation-based works [19, 11] use multi-modal distillation. Gao *et al*. [19] use multi-modal distillation from a video model to an image-audio model for action recognition. Chen *et al*. [11] propose to distill knowledge from single modal image and audio networks to a video network for video classification. Since the video network only inputs sequence of images without audio, multimodal fusion is not

required. Besides these existing works, Wang *et al*. [59] investigate that naïve approaches may not be the optimal solution in training multimodal classification networks due to inherent modality bias. They propose to use joint training by adding two separate uni-modal branches with weighted blending. Our learning mechanism is similar to this approach in terms of multi-task joint training but our training scheme is applied to multiple event-specific layers to address proper multisensory integration.

Due to the rising popularity of transformers [57], recent works [37, 46] design their transformer architectures with audio and visual signal inputs. In contrast to these heavy transformer based approaches, our work considers the multiple types of cross-modal interactions with light backbone networks, and applies to in-depth video understanding beyond simple recognition tasks.

Broader Audio-Visual Learning Tasks. Recent works on audio-visual learning use the natural correspondence between auditory and visual signals on different tasks than representation learning and action recognition, including audio-visual sound separation [13, 14, 15, 17, 18, 1, 69, 66, 67, 63, 56], sound source localization [47, 5, 48, 55, 23], audio generation [70, 35, 16, 68, 64] and audio-visual event localization [30, 53, 61]. Different from all these works, we focus on incorporating audio and visual modalities for multisensory integration without the assumption that they are always correspondent.

Cognitive Science. Our design is motivated by the findings in the numerous biology, psychology, and cognitive science study on multisensory integration in the brain [51, 49, 50, 25, 38, 39, 9]. Basically, they show that full pairwise correspondence at all time-steps of the audio and visual signals is not optimal because these signals contain different relationships [25, 49, 50]; *i.e*., relying on a single mechanism of simple concatenation or global pooling only addresses limited cases. The evidence in these studies also show that the brain solves two problems in perception: 1) to bind or segregate the different sensory modalities depending on whether they originate from a common or separate events; 2) to devise ways to integrate them properly if they go together. These studies suggest the human brain uses different types of perceptual factors - such as temporal, spatial, semantic, and structural - while integrating different sensory signals. In our work, we take inspiration from these studies and formalize the multisensory binding *vs.* segregation by designing multisensory event-specific layers.

3. Approach

The goal of our model is to understand and predict an accurate label that represents a video from the perspective of each multisensory layer. Most of the existing works [31, 29, 26, 62, 59] use a clip level classifier that takes a short clip (1 or 2 sec.) and then computes video-level predictions

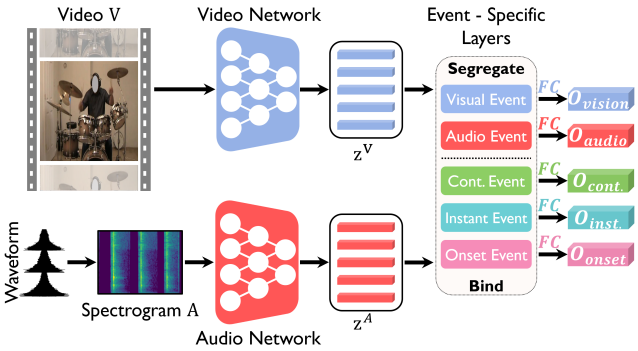

Figure 2. **Our multisensory framework.** The model consists of video and audio backbone networks that extract video-level features, $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$. The features are fed into *Event-Specific Layers* for multisensory integration. Each layer processes the features individually and predicts a category label.

by averaging the classification scores of each clip. These clip classifiers are learned by leveraging naïvely fused (*e.g*., concatenation followed by simple averaging) audio-visual features with the assumption that audio and visual signals are correlated and temporally aligned.

As aforementioned in Section 1, the existing approaches for video classification and understanding might be improved by considering more complex associations. First, audio and visual events in a video may not occur with a close association all the time. They can occur separately in each individual modality as well. Second, these audio-visual correspondences can have different characteristics such as continuous, rhythmic, or isolated instant events [49, 51]. Our proposed architecture addresses these concerns by using various multisensory event-specific layers.

Backbone Networks. Given a video clip $\mathbf{V}$ with its corresponding audio $\mathbf{A}$, our backbone networks extract features for each modality. We use a two-stream architecture, that leverages each modality separately, similar to other existing audio-visual learning works. Our backbone networks take an entire 10 sec. video and audio frames and extract features per-frame for each modality. We use a manageable size architecture, MCx, as a spatio-temporal video stream backbone by following [54] for extensive experiments. It takes a video $\mathbf{V}$ of T frames as input and generates a video embedding $\mathbf{z}^{\mathbf{V}}$ with dimensions $T \times D$. Our audio stream backbone is a modified version of the audio network used in [1]. A minimal modification, *e.g*., different kernel and stride sizes for a few layers, is applied to the audio network to make the temporal dimension of audio embedding and visual embedding the same. Our audio stream backbone take the log-mel spectrogram $\mathbf{A}$ of $10T$ frames and extract an audio embedding $\mathbf{z}^{\mathbf{A}}$ with dimensions $T \times D$ similar to video features. Thus, there is a corresponding audio feature for every video feature and we do not need any replication or tile operations to match audio and video feature dimensions.

3.1. Multisensory Event-Specific Layers

To deal with different multi-modal event types, we design expert layers as multi-heads of the audio-visual network (see Fig. 2). Defining i as the index of each layer, the layer takes $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$ from the backbone networks and outputs video-level prediction O_i. We explain each layer in detail below. Note that all the presented layers are parameter-free by itself.

Continuous Event Layer. This layer is the common integration method of audio and visual signals by performing temporal aggregation to each set of frame features from both modalities with the assumption of audio and visual signals are temporally correlated and aligned throughout a video. This temporal congruence between audio and visual signals play a key role for audio-visual sensory integration not only in cognitive science [49, 50] but also in audio-visual learning works [40, 53, 28, 21] as a dominant paradigm. The integrated audio-visual feature $\mathbf{z}_{cont.}$ is computed as follows:

$$\mathbf{z}_{cont.} = \frac{1}{T}\sum_{t=1}^{T} \texttt{concat}(\mathbf{z}_t^V, \mathbf{z}_t^A), \tag{1}$$

where $\texttt{concat}(\cdot)$ denotes the concatenation of two vectors and t the video time step. The continuous layer feature $\mathbf{z}_{cont.}$ is obtained by temporal aggregation over all time steps T by average pooling.

Instant Event Layer. Another type of audio-visual events that frequently occur is sparse and isolated instant ones. These interesting actions happen when both audio and visual signals are semantically correlated and synced for a short time as a few important moments rather than a long temporal duration. The assigned task for this layer is performed by finding the time steps (moments) that have the highest correlation scores between audio and visual features, $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$ respectively. Figure 3 shows that the moments with the highest scores are located only in the last part of the video where the car appears in the scene and it is correlated with the car sound (visualized as colored frames). This provides well-associated moments between audio and visual events. The remaining parts of the video are not useful for audio-visual integration as it only shows an empty road.

To find such moments, correlation scores are computed by pairwise dot products between audio and visual embeddings [21, 1] at the same time step, then the scores are used to compute audio-visual feature $\mathbf{z}_{inst.}$ as follows:

$$\mathbf{z}_{inst.} = \frac{1}{|\mathcal{K}|}\sum_{t\in\mathcal{K}} \texttt{concat}(\mathbf{z}_t^V, \mathbf{z}_t^A), \tag{2}$$

where $\mathcal{K}$ denotes a set of the top-k time steps according to the high correlation scores as $\mathcal{K} = \texttt{top-k}(\mathbf{S}_{av})$, and the score list $\mathbf{S}_{av}[t] = \mathbf{z}_t^V \cdot \mathbf{z}_t^A$. That is, the instant layer feature is obtained by averaging the features at the top-k time steps.

Onset Event Layer. Another type of audio-visual event can be integrated on event occurrences at regular points in time, *i.e.*, rhythm [51, 9]. For example, sounds occur rhythmically and repetitively in dancing, musical instruments, and bird-calling events as they have a prominent property in audio modality aligned with visual signals. The onset event layer is designed to leverage audio onsets which give information about rhythms and beats [12, 58], musical notes, and as well as the beginning of audio events [43, 27]. In Figure 3, the visual event (typebar hits the screen) occurs at the same time as the onset moments (pink-colored dots). Furthermore, almost equal time gaps between the onset moments show that this event is rhythmic.

We compute $\mathbf{z}_{onset}$ as follows:

$$\mathbf{z}_{onset} = \frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}} \texttt{concat}(\mathbf{z}_t^V, \mathbf{z}_t^A), \tag{3}$$

where $\mathcal{O} = \texttt{onset}(\mathbf{A})$ denotes audio onset moments. We used audio for computing onset moments, because computing the audio onset is efficient and distinctive compared to that of complex visual data. We can simply implement $\texttt{onset}(\cdot)$, *e.g.*, by measuring magnitudes of audio signal, but for better onset localization, we use the standard audio libraries [33] for computing the audio onset. This returns a set of time indices that onsets exist in the range of $\{1, \cdots, T\}$.

Visual Event Layer. Until now, our multisensory layers are inspired by the human cognitive ability for multisensory integration as binding the multimodal signals if they are correlated and separating them otherwise. Considering some actions are soundless ("handshakes", "stretching leg", *etc.*) or some scenes have irrelevant sounds, integrating these irrelevant sound signals to visual features acts as a detrimental outlier. Thus, the visual event layer is designed to recognize the events only from the visual perspective. It performs the task of assigning zero-valued audio features for each visual frame feature and applying Eq. (1) to output $\mathbf{z}_{visual}$.

Audio Event Layer. Analogously to the visual event layer, scenes might have events that are outside of the field of view but still hearable or the visual signals may be completely unrelated to the accompanying audio. Additionally, some videos might have poor visual signals. To make our network use audio modality only, the audio event layer assigns zero-valued visual features to each audio frame feature and applies Eq. (1) to compute $\mathbf{z}_{audio}$.

3.2. Training

With the backbone and the event-specific layers, we obtain different representations from each layer with given identical inputs, *i.e.*, audio and visual features from the backbone networks. To make each layer produce a final C-class prediction output O_i, letting i be the index of each layer, separate fully-connected layers are used as in Figure 2. We train the whole network with the multi-heads in the multi-task joint learning manner. We impose the same loss to the individual layers with a supervisory label, where the same

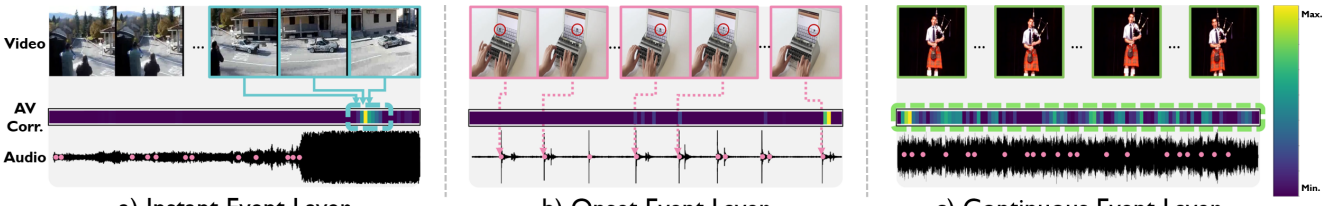

Figure 3. **Sampling position of audio-visual event layers.** We show how audio-visual event-specific layers perform their time index-based pooling operation. a) *Instant Event Layer* picks the moments where audio and visual features are highly associated, highlighted by the AV Correspondence heatmap in the middle row. b) *Onset Event Layer* only pools audio onset moments (pink-colored dots on the waveform) into the feature computation. c) *Continuous Event Layer* adopts the traditional global average pooling with uniform sampling to obtain global context information. Pink-colored dots indicate audio onset moments and are visualized just for reference purpose in (a) and (c).

single label is given across the heads, as:

$$\mathcal{L}_{multi} = \sum\nolimits_{i \in \mathcal{E}} \mathcal{L}_i(O_i, y), \quad \text{where } O_i = \texttt{FC}_i(\mathbf{z}_i), \quad (4)$$

$\mathcal{E} = \{cont., inst., onset, visual, audio\}$, $\mathcal{L}(\cdot)$ is the cross entropy loss, $\texttt{FC}(\cdot)$ is the fully-connected layer, O and y are the prediction output and label, respectively. We equally weight each loss.

The imposed losses could appear redundant but it has been shown to be effective in the previous multimodal learning study [59], where dominance to a specific modality head can be balanced by this similar approach. In our case, we balance across event types by encouraging to possess supervision relevant signals as much as possible.

4. Experiments

We first evaluate our method for video-level classification on four audio-visual datasets. Then, we show additional weakly-supervised features of our proposed method: the capability of multi-label prediction from the single-label training and the sound source localization task without any additional training. We also analyze characteristics of the event-specific layers. Last, we show that the proposed layers enables an event-characteristic analysis of existing datasets that may connect to a number of potential applications.

4.1. Setup

Datasets. We experiment our method on five video datasets: **VGGSound** [10] and **Kinetics-400** [24] for action recognition datasets, and the former is specifically designed for audio-visual learning. **Kinetics-Sound** [4] is a subset of Kinetics sub-sampled for audio-visual learning tasks, **AVE** [53] is for audio-visual event localization and **LLP** [52] is a multi-label dataset for audio-visual video parsing.

Implementation Details. More details can be found in the supplementary material. We follow the prior arts [1, 62] for audio preprocessing. For all the experiments, we sample audios with 16kHz sampling rate and all the input audio length is trimmed to 10 seconds. We transform the audios to log-mel spectrograms with size of 1000×80, and we use a modified audio network from [1]. MC3-18 [54] is used as the video network and it takes $T = 100$ frames of size 112×112 as input. We set $|\mathcal{K}| = 10$ for the instant event layer computation.

Dataset	Audio Only	Vision Only	Naïve AV	Ours
VGGSound	47.0	40.9	57.1	**59.1**
Kinetics-Sound	64.2	80.5	86.1	**88.3**
AVE	79.1	76.1	86.0	**87.8**
Kinetics	21.4	61.0	66.6	**67.0**

Table 1. **Video-level classification performance of our proposed model and baselines.**

We apply the same training process for each dataset as follows. First, we train the audio backbone network from scratch with a given target dataset. The video backbone network is initialized by using MC3-18 pre-trained on Kinetics-400 and fine-tuned on the target dataset. Last, we train our multi-task model with the event-specific layers in the end-to-end manner by using these pre-trained backbone networks as initialization.

4.2. Analyses on Video Understanding Tasks

Effectiveness on Video-Level Classification Video classification is a task to classify a video by a single label. Since our model outputs multiple predictions from event-specific layers, we integrate them by majority voting to output a single prediction as $\mathcal{P}_{vote} = \arg\max_k \sum_i I(p_i{=}k)$, where p_i is a predicted label from the i^{th} event-specific layer defined as $\arg\max_j O_{ij}$, and j is an index of the vector O_i, I is an indicator function returning 1 for true statement and 0 otherwise. In case of disagreement among the layers that no majority consensus exists, the label from the most confident layer is selected.

We conduct a series of experiments to show how well our model predicts video-level labels. We compare the performance of our model with baselines on different datasets in Table 1. Note that our goal here is *not* to compete on classification accuracy with any other expensive video recognition models. Rather, we show that our event-specific layers analyze videos from distinctive perspectives in terms of modalities and event characteristics, which leads to improvement

Figure 4. **Single- to multi-label prediction on VGGSound.** The original annotations are single labels, whereas typical videos contain multiple events, actions, or categories. Our multi-head design predicts multi-labels that enable a more comprehensive description of videos.

in classification.

Accuracies on the uni-modal networks in Table 1 depict the accuracy of the backbone networks trained with single-stream modalities. The naïve audio-visual model (*Naïve AV*) represents audio-visual networks that leverage the late fusion approach (simple concat. and global pooling) for final representation as used in the prior works [59, 28, 31]. As shown in Table 1, our approach offers improvement to overall performance in the benchmark datasets. Our model is more effective on the datasets that are designed with audio-visual correspondence, *e.g.*, VGGSound, Kinetics-Sound, and AVE, with the improvements around 2%. Our performance improvement is less significant on Kinetics, which is consistent with [37], since it is a visually-dominant dataset, where many videos' sounds are not correlated to visual signals (We further analyze the datasets in the "**Revealing event-characteristics of a dataset**" paragraph).

Does our multi-head design have multi-label prediction capability? Typically, large-scale video datasets, *e.g.*, VGG-Sound and Kinetics, are annotated with single labels of dominant events. Therefore, the annotation ignores the other events that may co-occur in a video. Our network consists of multiple event-specific layers and each layer outputs its own label prediction. As a simple way to extract multiple outputs, we collect each layer's most confident predictions that directly form a multi-label prediction set. Specifically, let p_i be a predicted label from the i^{th} event-specific layer. Then, a set of label predictions $\mathcal{P}$ can be defined as $\mathcal{P} = \bigcup_i \left(\arg\max_j O_{i,j}\right)$. In this simple way of aggregating multiple predictions, we can analyze if the event-specific layers enable us to see the contents of a video from different perspectives.

This multi-label analysis may be used to answer the questions of "Are the multi-label predictions of our network correct?" or "Does an existing dataset like VGGSound contain multiple events in a video but only annotated with a single label?". To answer the first question, in Table 2, we conduct an experiment on the LLP dataset as it contains multi-labels per video (The average number of labels per video is 1.81), which allows to evaluate the correctness of the multi-labels predicted by our network. We train our model on LLP, but with the restriction of a single label per video, so that we can measure the ability of our multi-head layers for predicting correct multi-labels despite the single-label training.

Dataset	Ours	Naïve AV				
		Top-1	Top-2	Top-3	Top-4	Top-5
LLP	**0.72**	0.66	0.70	0.63	0.56	0.50

Table 2. **Multi-label prediction measured by F1 score.**

We compare our results with top-K results of the Naïve-AV model as a baseline. The top-K naïve approach outputs K number of predictions, while our model outputs multi-label predictions dynamically depending on the consensus of the event-specific layers.

In Table 2, we show that our method has indeed a notably better capability to predict correct multi-labels over Top-K baselines, despite training only with single-label per a video clip. In addition, it shows a favorable feature of our multi-label prediction. Although the Top-2 performance is comparable to ours, our method can adaptively decide the number of output labels (K), while the baseline Top-K cannot decide what K to use in practice because the number of ground truth multi-labels is unknown in advance. For example, the video shown in the 2^{nd} row of Figure 4 contains the people playing more than two instruments. In this case, Top-2 only outputs two predictions while our model outputs more than two predictions. On the other hand, when there is only one dominant event, our model tends to output only one prediction while the Top-2 baseline enforces to still output two predictions.

To answer the second questions, "Does an existing dataset like VGGSound contain multiple events in a video but only annotated with a single label?", we additionally conduct multi-label check evaluation. However, there is no multi-label ground truth for the VGGSound dataset. Instead, we check how many of the total predictions from our network actually do match with human answers according to given video contents. For this test, we sample a partial subset of near 1200 videos from the VGGSound dataset and ask 12 subjects to evaluate predicted labels obtained by our model. When our network is evaluated on the VGGSound dataset, the empirical cardinality of $\mathcal{P}$, *i.e.*, the average number of different predicted labels, is 2.21.[3] The user study shows that 62% of all predicted labels match with human selections, which means our network outputs 1.4 correct labels per sample on average. Figure 4 qualitatively shows multi-

[3]Since we utilize five types of layers, the cardinality of the prediction set is in $1 \leq |\mathcal{P}| \leq 5$.

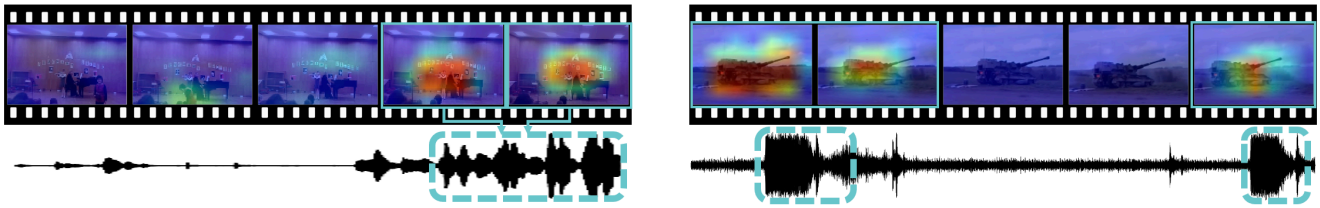

Figure 5. **Sound source localization.** Our backbone networks correctly localize the sound spatially and time-wise as a natural outcome of the model without any explicit training for sound localization.

Dataset	Continuous	Instant	Onset	(Ins.∨Ons.) ∧ (¬Cont.)
VGGSound	354	407	230	778
Kinetics-Sound	18	18	9	34
AVE	7	4	3	12
LLP	25	47	22	74
Kinetics	366	567	258	985

Table 3. **Layer-wise statistics of correctly predicted videos.**

labels align with human subjects and our network. This study would evidence that *VGGSound is indeed a multi-label dataset, and a single label is insufficient to describe the videos properly*. In this way, we annotate a partial subset of near 1200 videos in the VGGSound dataset, called Multi-labeled VGGSound. The details are in the supp. material.

Are our learned features interpretable? Our event specific layers are designed to differently select audio-visual correlated moments: *e.g.*, the instant event layer catches highly correlated audio-visual moments. Thus, to gain a better understanding of these moments and analysis, we visualize sound localization responses $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = \mathcal{V}_t \cdot \mathbf{z}_t^A$, $\mathcal{V}_t \in \mathbb{R}^{H' \times W' \times D}$ is the visual activation from the last convolution layer of video backbone network and $\mathbf{z}_t^A \in \mathbb{R}^D$ is the audio embedding at moment t [48, 1]. Note that this does not require any separate additional training.

We qualitatively show in Figure 5 that the features from our backbone networks can plausibly locate a sound source despite no separate training for this task. The localization response only activates when the girl plays flute, otherwise inactivated. This confirms that our model not only attends *where* the source appears in the video spatially but also attends *when* the event sound occurs time-wise. Refer to the supp. material for more results. Thus, our learned representation is well-trained such that it sufficiently localizes sources of events.

Are the event-specific layers complementary? To show this, in the first three columns on Table 3, we first count unique video samples that one of the continuous, instant, and onset layers classifies into true classes while the rest two layers predict different classes. Interestingly, the numbers of those samples for each layer are comparable. This shows that each layer indeed looks for different audio-visual characteristics in videos.

In the last column, we count video samples that the continuous layer (the conventional fusion layer) fails but the instant or onset layers predict correctly. It shows that the instant and onset layers capture a significant amount of correct samples compared to the samples uniquely captured by the continuous layer. That is, our new proposed layers, the instant and onset layers, contribute noticeably over the continuous one, and are indeed complementary.

For typical events in videos, the number of informative moments (features) may be less than uninformative ones for some videos, as in the "car" example of Figure 3. Thus, using an informative subset of time step features, *i.e.*, instant or onset, may improve the accuracy for these videos since irrelevant features are ignored. Furthermore, the difference between the instant and onset event layers can be also seen in the same "car" example. The onset event layer uses the onset moments, *i.e.*, pink dots, that are grouped in the part before the real action starts whereas the instant event layer captures the instant, highly audio-visual correlated moments of the video, *i.e.*, blue dotted box.

In contrast, the continuous event layer pools information of both modalities' features from every time step without consideration of their correspondence. Clearly, a single type layer would not be the best way for multi-modal fusion, because audio and visual events have different relationships. We show that our event-specific layers are complementary each other, which comes from selection mechanisms of moments by design (refer to Eqs. 1, 2 and 3). Thereby, we show our model enables a more in-depth interpretation of videos.

Layer Visualizations. To have a better insight on what each event-specific layer learned, we visualize some of the videos that are maximally activated by each audio-visual layer in Figure 6. The instant event layer has short period and high intensity-like patterns; the onset event layer captures rhythmic-like patterns; and the continuous event layer shows temporally constant-like events. We also visualize the prediction of each layer. The event-specific layer that corresponds to the event type of a video outputs a correct prediction while the remaining event-specific layers fail.

Analysis on the event-characteristics of datasets. We apply our method to understanding the event-characteristics of datasets. Each dataset has different event properties such as a large portion of the videos may contain a specific event type dominantly [60], *i.e.*, Kinetics is a visually biased dataset [37]. We pose the problem as finding the most dominant event-type in a given dataset by analyzing every video. Our method can easily detect the dominant event type

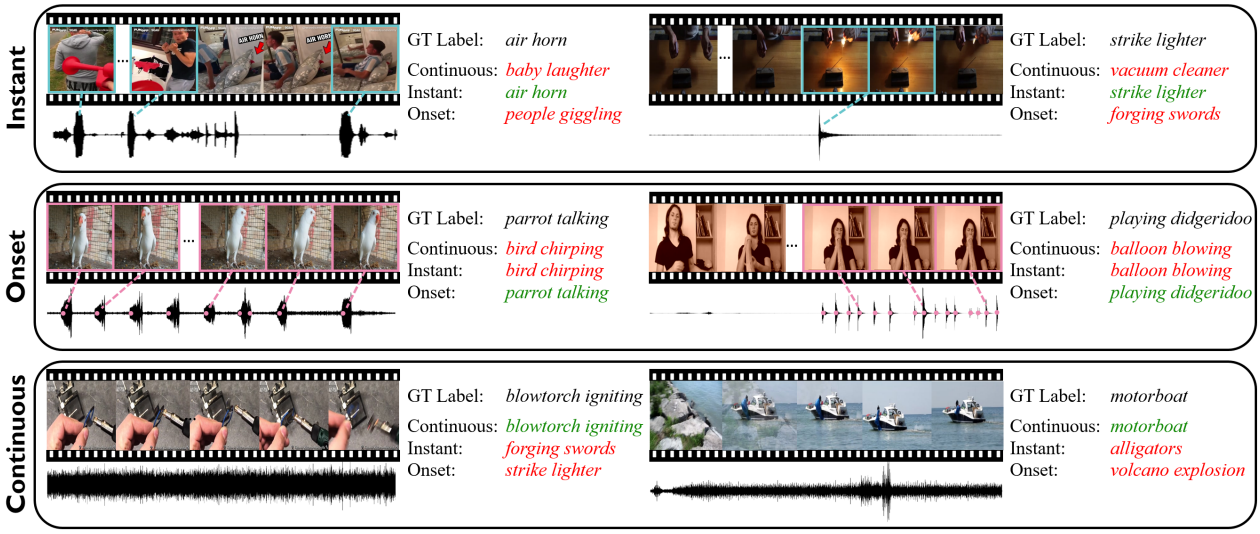

Figure 6. **Visualization of the representative characteristics from the event-specific layers.** Each layer captures a distinctive audio-visual characteristic. Note that our multisensory model not only detects the event types correctly but also makes an accurate category prediction within the layers.

Dataset	Continuous	Instant	Onset	Visual	Audio	Total
VGGSound	62	39	68	41	**99**	309
Kinetics-Sound	2	5	**13**	11	0	31
AVE	**10**	4	9	1	4	28
LLP	4	5	6	3	**7**	25
Kinetics	73	89	69	**167**	2	400

Table 4. **Event-characteristics of datasets.** We report the number of categories asssigned to each multisensory layer.

of each video by checking which event-specific layer has the highest score on the ground truth class y as $\arg\max_i O_{i,y}$.

With this technique, we find the event-characteristics per category and per dataset. For finding categories with dominant event type, we apply the majority voting rule to each category and assign the most voted event-type label to the category. In this way, we can obtain the event-type characteristics of the categories in the dataset. In our test, categories such as "bowling impact" or "splashing water" are associated with the instant layer, or "air conditioning noise" is assigned to the continuous layer. See the supplementary material for the category-wise event-type assignment results. Table 4 shows the summary of the number of categories that are assigned to each layer for the datasets. Our analysis shows consistent results with the prior knowledge about these datasets. The results clearly show that Kinetics is visually dominant as the number of categories assigned to the visual layer is the highest. AVE is curated for audio-visual learning and our method validates it by the dominance of the AV layers. LLP [52] reports that the majority of the annotated events are audio events. Our analysis also confirms a tendency to the audio modality.

Additionally, we perform an experiment to see how many categories of Kinetics-Sound match with the audio-visual categories that our method found in Kinetics. This reveals that 66% of the Kinetics-Sound categories are matched. Thus, our event-type selection gives consistent results with human selections. Please see the supp. material for details.

5. Concluding Remarks

We present a multisensory model with event-specific layers that incorporates different audio-visual relationships and demonstrate the efficacy of our model on five different video datasets with a diverse set of videos. Unlike the prior audio-visual models, our event-specific layers output multiple predictions. This leads to new future research directions for audio-visual understanding. We conclude with discussion about potential applications of our work followed by limitations in the supplementary material.

Potential Applications. Our method can open useful potential applications — 1) Modality-level video understanding: Within a single video clip, different modalities might play a key role at different timestamps in video understanding. Our method can tell which modality to rely on to understand ongoing events in a video using confidence from each layer. This kind of modality-level video understanding, as opposed to class prediction, would be crucial and complementary to existing methods, 2) Missing label detection: Also, by virtue of multi-label prediction property of our method, our method can discover potential labels and therefore be used to build up a more comprehensive dataset by detecting missing labels, and 3) Dataset Retargeting / Cleanup: Our method can be further used to retarget or clean-up by modality/event-level classification for each video, so that we can easily create an application-specific sub-datasets. These would improve the video annotation system when being integrated into it. More detail scenarios are discussed in the supplementary material.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision (ECCV)*, 2020.

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision (ECCV)*, 2018.

[6] Yuki M. Asano, Patrick Mandela, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[9] Andrew Bremner, David Lewkowicz, and Charles Spence. *Multisensory Development*. Oxford University Press, 2012.

[10] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[11] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. In *ACM Transactions on Graphics (SIGGRAPH)*, 2018.

[13] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (SIGGRAPH)*, 2018.

[14] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. *European Conference on Computer Vision (ECCV)*, 2018.

[16] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[18] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*, 2018.

[21] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[22] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[25] Christoph Kayser and Ladan Shams. Multisensory causal inference in the brain. *PLoS Biol*, 13(2):e1002075, 2015.

[26] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[27] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D Plumbley. Sound event detection and time-frequency segmentation from weakly labelled data. *arXiv preprint arXiv:1804.04715*, 2018.

[28] Bruno Korbar, Du Tran, and Lorenzo Torressani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[29] Bruno Korbar, Du Tran, and Lorenzo Torressani. Scsampler: Sampling salient clips from video for efficient action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[30] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[31] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local video representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[33] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[34] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[35] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[36] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[37] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[38] Olha Nahorna, Frederic Berthommier, and Jean L. Schwartz. Binding and unbinding the auditory and visual streams in the mcgurk effect. *The Journal of the Acoustical Society of America*, 132:1061–1077, 2012.

[39] Olha Nahorna, Frederic Berthommier, and Jean L. Schwartz. Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the mcgurk effect. *The Journal of the Acoustical Society of America*, 137:362–277, 2015.

[40] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, 2018.

[41] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[42] Andrew Owens, Jiajun Wu, Josh McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision (ECCV)*, 2016.

[43] Arjun Pankajakshan, Helen L. Bear, and Emmanouil Benetos. Onsets, activity, and events: A multi-task approach for polyphonic sound event modelling. *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE '19), New York, NY, USA*, 2019.

[44] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[45] Mandela Patrick, Po-Yao Huang, Ishan Misra, Florian Metze, Andrea Vedaldi, Yuki M Asano, and João F Henriques. Space-time crop & attend: Improving cross-modal video representation learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[46] Merey Ramazanova, Victor Escorcia, Fabian Caba Heilbron, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context. *arXiv preprint arXiv:2202.04947*, 2022.

[47] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes: Analysis and applications. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. To appear.

[49] Charles Spence. Audiovisual multisensory integration. *Acoustical Science and Technology*, 28(2):61–70, 2007.

[50] Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception and Psychophysics*, 73:971–95, 05 2011.

[51] Yi H. Su. Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams. *Frontiers in Integrative Neuroscience*, 8, 2014.

[52] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision (ECCV)*, 2020.

[53] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, 2018.

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

[55] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[56] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel P. W. Ellis, and John R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *International Conference on Learning Representations (ICLR)*, 2021.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[58] Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2020.

[59] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[60] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. In *British Machine Vision Conference (BMVC)*, 2020.

[61] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[62] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[63] Xudong Xu, Bo Dai, and Lin Dahua. Recursive visual sound separation using minus-plus net. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[64] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[65] Yunhua Zhang, Ling Shao, and Cees G. M. Snoek. Repetitive activity counting by sight and sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[66] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[67] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European Conference on Computer Vision (ECCV)*, 2018.

[68] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[69] Hang Zhou, Xudong Xu, Lin Dahua, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision (ECCV)*, 2020.

[70] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.