Decompose the Sounds and Pixels, Recompose the Events

Varshanth R. Rao [1], **Md Ibrahim Khalil** [1,2], **Haoda Li** [1,3], **Peng Dai** [1 *], **Juwei Lu** [1]

[1]Huawei Noah's Ark Lab
[2]University of Waterloo, Canada
[3]University of Toronto, Canada
{varshanth.rao1, md.ibrahim.khalil, haoda.li, peng.dai, juwei.lu}@huawei.com

Abstract

In this paper, we propose a framework centering around a novel architecture called the Event Decomposition Recomposition Network (EDRNet) to tackle the Audio-Visual Event (AVE) localization problem in the supervised and weakly supervised settings. AVEs in the real world exhibit common unravelling patterns (termed as *Event Progress Checkpoints* (EPC)), which humans can perceive through the cooperation of their auditory and visual senses. Unlike earlier methods which attempt to recognize entire event sequences, the EDRNet models EPCs and inter-EPC relationships using stacked temporal convolutions. Based on the postulation that EPC representations are theoretically consistent for an event category, we introduce the State Machine Based Video Fusion, a novel augmentation technique that blends source videos using different EPC template sequences. Additionally, we design a new loss function called the Land-Shore-Sea loss to compactify continuous foreground and background representations. Lastly, to alleviate the issue of confusing events during weak supervision, we propose a prediction stabilization method called Bag to Instance Label Correction. Experiments on the AVE dataset show that our collective framework outperforms the state-of-the-art by a sizable margin.

Introduction

As videos transform the landscape of information exchange and entertainment, audio-visual understanding becomes vital to augment the related applications. True to the quote by Harry Houdini - *"What the eyes see and the ears hear, the mind believes"*, the successful perception of videos is powered through the activation and interplay of the visual and auditory sensory streams. A pragmatic testimony of the necessity of the audio-visual integration is the McGurk effect (McGurk and MacDonald 1976) in speech perception (Schwartz, Berthommier, and Savariaux 2002). Additionally, ambient noise in real life like traffic or wind may mask or overlap with the sound of interest, rendering audio to be the noisier modality. The marriage of the two modalities has benefited various audio-visual tasks such as sound source localization and separation (Zhao et al. 2018; Owens and Efros 2018; Korbar, Tran, and Torresani 2018; Arandjelovic and Zisserman 2018; Senocak et al. 2018), audio-visual speech recognition (Noda et al. 2015; Afouras et al. 2018), etc.

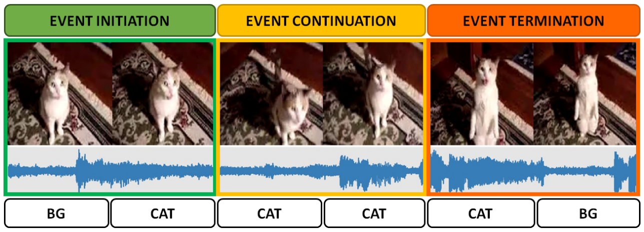


Figure 1: A depiction of event atomization. Events can be decomposed into Event Progress Checkpoints (top row) according to the temporal position along the event line. The corresponding audio-visual sequence is displayed in the middle row, while the ground truth segment labels are indicated in the bottom row.

Under the umbrella of video understanding, the Audio-Visual Event Localization (AVEL) problem (Tian et al. 2018) has garnered increasing attention. AVEL entails the localization and identification of events which are both audible and visible. Supervised Event Localization (SEL) is to perform AVEL when the segment-level labels are available for supervision while for the Weakly Supervised Event Localization (WSEL) task, only the video level labels are available. Although the accompanying AVE dataset (Tian et al. 2018) is size-wise small (4143 videos covering 28 event categories), it is currently the only dataset to offer segment level audio-visual labels, hence facilitating the SEL task.

Previous works resolve short and long-term temporal relationships by sequentially traversing through the video using memory networks like LSTMs. Digressing from their approach and inspired from the viewpoint of bottom-up temporal action localization (Lin et al. 2018, 2019; Liu et al. 2019), we propose that events can be atomized into three Event Progress Checkpoints (EPCs); Event Initiation, Event Continuation, and Event Termination. We visualize these EPC formations in Fig. 1, where member segments delineate the commencement, uninterrupted procession, and completion of the event respectively. A natural choice for aggregating segments into EPCs is through Temporal Convolutional Networks (TCN) (Lea et al. 2017). By stacking TCN layers, we can establish temporal dependencies (potentially longer

[*]Corresponding author.

than recurrent networks (Bai, Kolter, and Koltun 2018)) at different granularities. Thus, we coalesce the fundamental concepts of the TCN and U-Net (Ronneberger, Fischer, and Brox 2015) to create a novel architecture called the Event Decomposition Recomposition Network (EDRNet). Leveraging EPCs (in place of segments) as the foundation units can lead to better generalization since an EPC representation of a category should be applicable across all videos of that category. Based on this assumption, we propose a State Machine Based Video Fusion technique which expands the current dataset by blending conceptually similar videos using random state-machine generated EPC sequences.

Typically, AVEL is treated as a segment classification problem and supervision involves treating segments as independent units. As a consequence, the relationships within and between continuous foreground (FG) and background (BG) segments (called patches) are overlooked. Instead, we view patches as neighborhoods with internal (FG-FG/ BG-BG) similarities and external (FG-BG) differences. We strengthen these relationships by supervising the EDRNet with a new Land-Shore-Sea loss function.

Prior works fail to explain the underlying cause of the performance gap between SEL and WSEL methods. In our investigation, we observe that WSEL methods suffer from a higher confusion between similar AVEs (such as helicopter vs airplane) due to smaller error gradients arising from weak supervision. To alleviate this issue, we extend a Multi-Instance Learning (MIL) method termed as the Bag to Instance Label Correction to refine FG localization predictions inside confident neighborhoods. In summary, our contributions can be highlighted as follows:

1. Based on event atomization, we propose the Event Decomposition Recomposition Network (EDRNet), a novel TCN-based architecture to tackle the AVEL problem.
2. We describe how the State Machine Based Video Fusion blends conceptually similar but content-wise different videos from segment annotated source videos, thus expanding the AVE dataset and fueling data-hungry deep networks.
3. We describe a heuristic method, the Bag to Instance Label Correction, to rectify incorrect FG predictions amongst confident FG neighborhoods.
4. Experimental results show that the EDRNet framework outperforms the state-of-the-art methods for both SEL and WSEL tasks on the AVE dataset.

Related Works

Audio-Visual Event Localization (AVEL): Event localization entails the identification of regions of an input sequence corresponding to the annotated event(s). Event localization has been classically viewed as an MIL problem (Dietterich, Lathrop, and Lozano-Pérez 2001), where only the coarse level labels are available. Such works have explored the use of the visual modality in videos (Lai et al. 2014; Xu, Yang, and Hauptmann 2015), audio modality in sound files (Parascandolo, Huttunen, and Virtanen 2016; Adavanne et al. 2019), and combinations of both along with video attributes and text in videos (Lan et al. 2012; Oh et al. 2014). Recently released, the AVE dataset (Tian et al. 2018) contains videos annotated both at the video and segment level for audible and visible events. (Lin and Wang 2020) devise an Audiovisual Transformer to use audio as the guiding modality to refine visual features by performing spatial attention on contextual frames and instance attention to locate the sound-source within a frame. The Positive Sample Propagation module by (Zhou et al. 2021) calculates similarity matrices between audio and visual features of different segments and thresholds them to eliminate insignificant audio-visual pairs. These matrices are used to co-refine similar segments together before fusing the modality information and learning temporal dependencies using LSTMs. Different from prior methods, we capture temporal dependencies progressively by stacking temporal convolutions. Further, in our qualitative analysis, we demonstrate that cross-modal guidance can be achieved *implicitly* through modality fusion.

Multi-modal Fusion: Multi-modal learning can result in more robust inferences since each modality can supply unique information relevant to the downstream task. Adopting the correct fusion technique becomes vital for exploiting each modality. Although prior works across different domains (Snoek, Worring, and Smeulders 2005; Gunes and Piccardi 2005; Kim and Bansal 2019; Alberti et al. 2019) have explored and compared early and late stage fusion methods, it remains inconclusive which of the two is superior. In (Lan et al. 2012), double fusion is proposed wherein late fusion is achieved by aggregating results of classifiers which operate on combinations of early fused features. Hybrid fusion is described in (Atrey et al. 2010) wherein the outputs of early and late fusion units are further merged to yield the final decision unit. Past AVEL methods (Tian et al. 2018; Ramaswamy 2020; Zhou et al. 2021) adopt LSTMs to fuse the audio and visual modalities while learning temporal alignments. In our work, the EDRNet executes *dual-phase modality fusion* wherein an early fusion branch amasses crucial cross-modal dependencies and later incrementally refines individual modality representations.

Methodology

Problem Statement and Notations

In the AVEL problem, each video sequence is split into N non-overlapping segments. The segment level event label is denoted by $y_t = \{y_t^c | y_t^c \in \{0,1\}, \sum_{c=0}^{C-1} y_t^c = 1\}$ while the video level event label is denoted by $y = \{y^c | y^c \in \{0,1\}, \sum_{c=0}^{C-1} y^c = 1\}$. Here C denotes the number of event classes inclusive of a BG event indicating independently audible (or visible) events or the absence of an event. For each video segment, the audio and visual features are extracted and denoted as $\{f_t^A, f_t^V\}_{t=1}^{N}$ respectively. Here $f_t^A \in \mathcal{R}^{d_a}$ and $f_t^V \in \mathcal{R}^{d_v \times S}$ where d_a is the dimension of the audio features, d_v and S are the dimension and the spatial size of the visual feature maps respectively. Following (Tian et al. 2018), we fix the feature extractors and build our architecture on top of these local features. SEL and WSEL tasks entail the prediction of the segment level event label $\hat{y_t}$, wherein y_t is available to use for training in SEL and only the video level label y is available for WSEL.

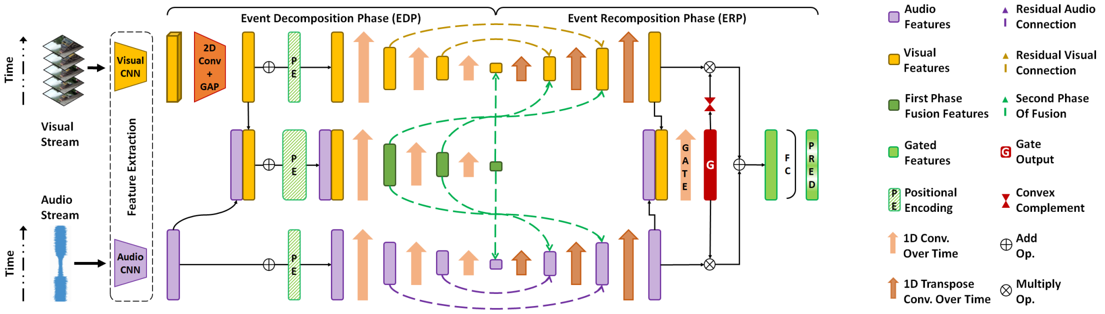


Figure 2: Overview of the Event Decomposition Recomposition Network. Modality-wise localization features are forged and refined in two phases: the Event Decomposition Phase (EDP) and the Event Recomposition Phase (ERP). The EDP summarizes the video into an event composition which the ERP leverages to effectively localize events in increasing temporal granularity. Consensus of the modalities is learned by a gating mechanism to yield the final audio-visual event localization predictions.

Event Decomposition Recomposition Network

The EDRNet, as depicted in Fig. 2, tackles both the SEL and WSEL tasks and operates in two phases; the Event Decomposition Phase (EDP) and the Event Recomposition Phase (ERP). The EDRNet utilizes the EDP to form a global understanding by encoding the video's event composition into temporally compressed audio, visual, and audio-visual representations. The audio and visual representations are then temporally upsampled in the ERP to "recompose" the event localizations from each modality's perspective at the segment level. Residual connections from the EDP supplies audio, visual, and audio-visual cues to the ERP's modality-wise localization branches. We detail both phases below.

Event Decomposition Phase: Given f_t^A and f_t^V, the EDP amasses information from fixed-sized segment sequences (decomposed events) into a global representation capable of describing the video's event composition. We extract the spatial information from f_t^V, using a 2D convolutional layer followed by a Global Average Pooling (GAP), yielding a condensed feature vector $\hat{f}_t^V \in \mathcal{R}^{d_v}$. To form composition perspectives, we build three independent modal branches; audio only, visual only and fused audio-visual, initialized as $D_A = \{f_t^A\}_{t=1}^N$, $D_V = \{\hat{f}_t^V\}_{t=1}^N$ and $D_{AV} = \{f_t^{AV}\}_{t=1}^N$ respectively where $f_t^{AV} = [f_t^A, \hat{f}_t^V]$ and $[.]$ is the concatenation operation. Following (Vaswani et al. 2017), we add positional encoding p_t to each branch's initial features $f_t^b \in \mathcal{R}^{d_b}$, $b \in \{A, V, AV\}$ to inform the subsequent network components about the positional (temporal) relationship between the local features within each branch. Hence we define p_t^i for all $i \in [1, d_b]$ and integer k as:

$$p_t^i = \begin{cases} sin(\omega_k t), & \text{if } i = 2k \\ cos(\omega_k t), & \text{if } i = 2k+1 \end{cases} \text{ where } \omega_k = \frac{1}{10^{8k/d_b}} \tag{1}$$

To learn event patterns at different temporal granularity, layers of temporal convolutions, which we term as decomposition operations, are employed in each modal branch. Specifically for each modal branch b, we define at layer l_{dec}, the output of a temporal convolution $\mathcal{F}$ with parameters $W_b^{l_{dec}}$ using kernel size k as:

$$D_b^{l_{dec}} = \mathcal{F}(D_b^{l_{dec}-1}, k; W_b^{l_{dec}}) \text{ where } D_b^0 = D_b \tag{2}$$

After L decomposition layers, we denote the modal branch outputs as $D_b^L \in \mathcal{R}^{N_{dec} \times d_L}$, where, N_{dec} and d_L are the temporal and feature map dimensions respectively.

Event Recomposition Phase: In the EDP, the audio-visual branch performs fusion of the constituent modalities to gather a joint perspective by forming cross-modal associations. EDP feature maps at layer l_{dec}, i.e., $D_b^{l_{dec}}$, contain both event specific positional information as well as event compositions local to the receptive field at layer l_{dec}. In the ERP, we use D_A^L and D_V^L as a foundation to forge progressive modality-wise localization branches using temporal transpose convolutions, which we term as recomposition operations. Additionally, we derive modality-specific guidance and dissemination of cross-modal knowledge through residual connections from the corresponding EDP's modality-specific layers and fusion layers at the same receptive field RF. The cross-modal knowledge gained by early fusion in the EDP and its flow into the ERP is termed as *dual-phase modality fusion*. It is noteworthy to highlight that without the inclusion of the positional encoding in the EDP, the decomposition operations gradually results in the loss of the events' position information within a video, which can be crucial for the ERP to form effective localization features. We formulate the recomposition output of branch $b' \in \{A, V\}$ at layer l_{rec} using temporal transpose convolutions of kernel size k and parameters $W_{b'}^{l_{rec}}$ as:

$$R_{b'}^{l_{rec}} = \mathcal{F}^T(R_{in}, k; W_{b'}^{l_{rec}})$$
$$R_{in} = \begin{cases} D_{b'}^L + D_{AV}^L, & \text{if } l_{rec} = 1, \\ R_{b'}^{l_{rec}-1} + D_{b'}^{l_{dec}} + D_{AV}^{l_{dec}}, & \text{otherwise} \end{cases} \tag{3}$$

where $RF(l_{rec}) = RF(l_{dec})$. After L layers of recomposition operations, we denote the recomposed modal branch outputs as $R_{b'}^L \in \mathcal{R}^{N \times d'_L}$. Here, R_A^L and R_V^L are features which respectively form the audio and visual perspectives of the event localizations. To successfully localize and identify AVEs, a consensus is required from both the modalities, the extent of which is context dependent. For example, to recognize a dog barking when its mouth is not clearly visible (weak visual signal), the model must seek out a barking sound in the audio modality. On the other hand, when its face is clearly visible but it is growling silently, the visual cue of the dog's slightly open mouth could aid in identifying the barking event. To capture this intricate cross-modal dependency, we learn a gating function through a temporal convolutional layer $\mathcal{F}_G$ with a sigmoid activation that operates on a fusion of both event localization perspectives.

$$G = \sigma(\mathcal{F}_G([R^L_A, R^L_V], k; W_G)) \in \mathcal{R}^{N \times d'_L} \quad (4)$$

where $[.]$ is the concatenation operation. We express the audio-visual perspective R^G_{AV} as a convex combination of the audio and visual localization perspectives with the coefficients as the gated output G and its complement respectively. We transform R^G_{AV} into localization predictions over C categories using an FC followed by a softmax activation.

$$R^G_{AV} = G \odot R^L_A + (1 - G) \odot R^L_V \quad (5)$$

$$\hat{y}_t = \text{Softmax}(W_{seg} R^G_{AV} + b_{seg}) \quad (6)$$

where denotes $\odot$ the Hadamard Product. For the SEL task, we supervise the segment level predictions using the multi-class cross entropy loss $\mathcal{L}_{seg} = CE(y_t, \hat{y}_t)$. The WSEL is formulated as an MIL problem where the video and segment level labels represent the bag and instance labels respectively. We use MIL pooling to aggregate the segment predictions into a single video level prediction $\hat{y}$.

$$\hat{y} = \frac{1}{N}\sum_{t=1}^{N} \hat{y}_t, \; \hat{y} \in \mathcal{R}^C \quad (7)$$

Land-Shore-Sea Loss

While $\mathcal{L}_{seg}$ can guide the EDRNet to classify individual segments correctly, it cannot capture the positive correlation within continuous FG (or BG) segments (called patches) and the negative correlation between the FG and BG patches. Ideally, FG features of the same event type should be closer while being further away from BG features. However, FG segments may not be discretely distinguishable from BG ones due to the following obstacles:

O1 A segment's audio/visual/audio-visual cues may be weak (or strong) leading to its features to be undesirably closer to that of the BG (or FG) event.

O2 FG events at the border may not span the entirety of the segment length and hence may be annotated as BG.

Motivated from natural topography, we view the FG patches as "Lands", BG patches as "Seas" and the FG event borders as the "Shores". Intuitively, a patch's representation should be a generalization of its constituent segments. If a majority of the segments in the patch exhibit strong audio-visual cues toward the correct event, the patch representation could be a beacon for its constituents. The Land (or Sea) loss hones the features of a FG (or BG) segment exhibiting O1 by bringing it closer to that of the Land (or Sea). Similarly, at the shore where segments can exhibit O1 and O2, the Shore loss draws the features of the event border (Shore) nearer to that of the FG patch (Land) while driving them apart from that of the BG patch (Sea). We depict the mechanism of each loss in Fig. 3. Let the gated features R^G_{AV} be denoted as L, S and Sh for land, sea and shore respectively. For a video sample i, we denote N^i_l, N^i_s and N^i_{sh} as the number of lands, seas and shores respectively. The land and sea loss are achieved by minimizing the L_2 distance between the average feature representations of the first half and second half of the land and sea patch respectively:

$$\mathcal{L}^i_{land} = \frac{1}{N^i_l}\sum_{l=1}^{N^i_l} ||\overline{L^{1st}_l} - \overline{L^{2nd}_l}||_2 \quad (8)$$

$$\mathcal{L}^i_{sea} = \frac{1}{N^i_s}\sum_{s=1}^{N^i_s} ||\overline{S^{1st}_s} - \overline{S^{2nd}_s}||_2 \quad (9)$$

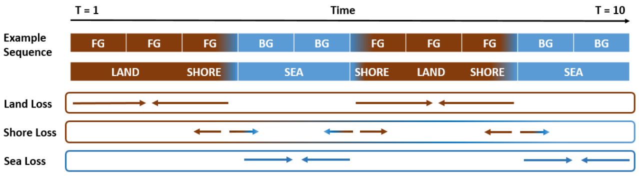


Figure 3: Visualization of the land, shore and sea losses. The land and sea losses aim to bring closer the features of continuous FG and BG segments respectively. The shore loss aims to pull closer the features of a FG border to that of neighboring FG segments (Land) while pushing it away from that of the neighboring BG segments (Sea).

The triplet loss is poised to implement the shore loss. The features of the shore represent the anchor sample while the average of the features of neighboring land and sea represent the positive and negative samples respectively:

$$\mathcal{L}^i_{shore} = \frac{1}{N^i_{sh}}\sum_{s'=1}^{N^i_{sh}} \Big[||Sh_{s'} - \overline{L_{s'}}||_2 - ||Sh_{s'} - \overline{S_{s'}}||_2 + \alpha\Big]_+ \quad (10)$$

where $[x]_+ = max(0, x)$ and α is the minimum margin to be maintained between $||Sh_{s'} - \overline{L_{s'}}||_2$ and $||Sh_{s'} - \overline{S_{s'}}||_2$. Thus, we define the LSS loss as $\mathcal{L}_{lss} = \mathcal{L}_{land} + \mathcal{L}_{sea} + \mathcal{L}_{shore}$ and train the EDRNet for the SEL task with the overall objective function as $\mathcal{L}_{SEL} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{lss}$

State Machine Based Video Fusion

We exhibited that events can be atomized into 3 Event Progress Checkpoints (EPCs): Event Initiation (EI), Continuation (EC), and Termination (ET). EPCs can be further dissected based on the minimum temporal length required for the EPC content to unfold. We denote BG sequences, EI, EC and ETs of length $L \in \{1, 2\}$ as BG_L, START_L, CONTINUE_L and END_L respectively. Table 1 contains the search and extract segment patterns for all $\langle$EPC/BG$\rangle$_L.

The samples extracted for an EPC from videos of a particular category should reflect similar semantic content. For example, the EI sequence for any baby crying usually involves the baby's eyes crunching up and the opening of the mouth, followed by the shrill sound of the cry. Thus, by replacing EPC-specific audio-visual content with other samples of the same EPC, we can synthesize an entirely new video expressing the same semantic event sequence. This allows localization models to focus on identifying the core conceptual

State	Segment Pattern	State Length
BG_1	[BG]	1 Segment
BG_2	[BG, BG]	2 Segments
START_1	[FG], Last N-1 Segments	1 Segment
START_2	[BG, FG]	2 Segments
CONTINUE_1	FG, [FG], FG	1 Segment
CONTINUE_2	FG, [FG, FG], FG	2 Segments
END_1	First N-1 Segments, [FG]	1 Segment
END_2	[FG, BG]	2 Segments

Table 1: Search & extraction segment patterns for all states. Within N-segment videos, we search for the complete segment pattern and extract only the ones enclosed within square brackets to form an instance for the state.

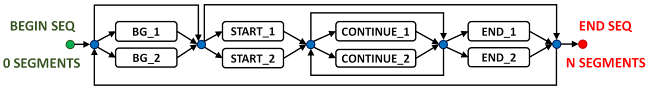


Figure 4: State machine used in the SMB Video Fusion technique. The state machine generates an N-segment state-sequence by following the displayed rules of state transition.

audio-visual correlations describing an EPC. However, this approach does not alter the sequence of EPCs and therefore limits the augmentation to the same sequence fingerprint. To introduce variation at the sequence level, we need to create different EPC blends. By considering all ⟨EPC/BG⟩_*L*s as states, we model the blend synthesis process using a state machine (SM) (shown in Fig. 4) since it must adhere to the positional constraints of EPCs.

Using Algorithm 1, we combine the above two strategies into a novel video augmentation technique called as the State Machine Based Video Fusion. Fig. 5 illustrates a snippet of an example output with the corresponding state sequence for the "Helicopter" event. As seen, the produced sequence is a coherently structured amalgamation of the source videos.

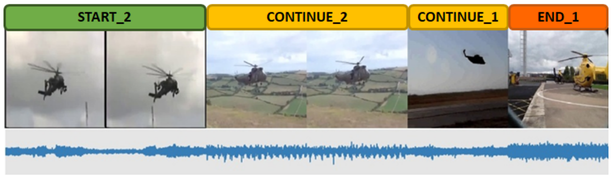


Figure 5: SMB fused video snippet of a "Helicopter" event. On top is the state sequence generated by the state machine. The bottom are the corresponding state-based clips which are stitched together from different helicopter videos.

```
Algorithm 1: SMB Intra-Class Video Fusion
Input: Training set videos of event type $e$ and length
       $N$, Number of output videos $N_v$
Output: $N_v$ fused videos of event type $e$
Initialize a database for each state
Identify the states for each video using Table 1
Store state specific video content into respective state
  databases
for $i \leftarrow 1$ to $N_v$ do
    Using the SM, generate a $N$-segment state
      sequence $SEQ_{state} = \{s_1, s_2, ..., s_{N_{state}}\}$
    Initialize output video $V_i$
    for $j \leftarrow 1$ to $N_{state}$ do
        Choose a random sample $v$ from $s_j$-database
        Append $v$ to $V_i$
    end
end
```

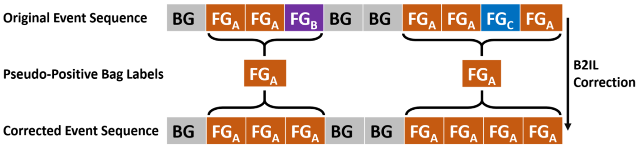


Figure 6: An example of B2ILC where incorrect $\mathrm{FG_B}$ and $\mathrm{FG_C}$ instances found in two pseudo-positive bags are corrected as $\mathrm{FG_A}$.

Bag to Instance Label Correction

From an MIL perspective of a localization sequence, a positive bag is usually of an FG event type, implying that negative instances can source from other FG types as well the BG. We tackle both sources separately by encapsulating continuous correct and incorrect FG predictions into a single pseudo-positive set. Discerning between BG and FG instances requires comprehension of video content which we entrust the model to perform. We can curtail the intra-FG event confusion within pseudo-positive sets by using the following MIL approach. In MIL, Witness Rate (WR) is the proportion of positive instances in the bag. When the WR is sufficiently high, the instance labels can be assumed to be that of the bag (Carbonneau et al. 2018). We term this process as "Bag to Instance Label Correction (B2ILC)". During inference time, we assume that a sufficiently trained localization model possesses a high FG (pseudo-positive label) precision, and hence we can treat continuous FG segment predictions to be a pseudo-positive bag. By setting the WR threshold as 0.5, the B2ILC becomes a strict majority voting process and on application to the pseudo-positive bag, all constituent instances (segments) would inherit the dominant FG label. We note that different from pure majority voting, B2ILC is constrained by the WR threshold. Fig. 6 demonstrates B2ILC with an example event sequence.

Experiments and Results

Dataset and Evaluation Metrics: The AVE dataset (Tian et al. 2018) is a subset of the AudioSet (Gemmeke et al. 2017) containing 4143 videos, each 10 seconds long, i.e., $N = 10$. There are 28 (+1 for BG) diverse event classes covering vehicle sounds, animal activity, instrument performances, etc. Video and segment level labels are available with clearly demarcated temporal boundaries. Events must be both audible and visible and spans for at least two seconds. We adopt the same train/validation/test split as (Tian et al. 2018). Recently, (Zhou et al. 2021) corrected the annotations for some test videos and report their performance on

AVEL Method	Dataset	WSEL Acc (%)	SEL Acc (%)
AVT (Lin and Wang 2020)	O-AVE	70.20	76.80
PSP (Zhou et al. 2021)	O-AVE	72.80	76.84
PSP (Zhou et al. 2021)	C-AVE	73.50	77.80
EDRNet (Ours)	**O-AVE**	**73.86**	**77.93**
EDRNet (Ours)	**C-AVE**	**74.57**	**78.91**

Table 2: Performance comparison with SoTAs (%) for the SEL and WSEL task on the O-AVE and C-AVE datasets.

their method on this corrected test set. We denote the AVE dataset with the *original* test set as **O-AVE**, and that with the *corrected* test set as **C-AVE**. Following prior works, we evaluate the localization performance using the global classification accuracy of segment level predictions.

Implementation Details: For a fair comparison with prior works, we utilize the same extracted audio and visual features (provided with the AVE dataset) using VGGish (Hershey et al. 2017) and VGG19 (Simonyan and Zisserman 2014) networks pretrained on AudioSet (Gemmeke et al. 2017) and ImageNet (Russakovsky et al. 2015) respectively. We configure the EDRNet with $k = 3$, $L = 4$, and a network width $d_l = 768$ for all layers. Sourcing the training set, we generate 250 samples per category using SMB video fusion. The optimization parameters for training EDRNet are specified in the *Supplementary Material*. Hyperparameter tuning was performed using the validation set.

Benchmarking against SoTAs

In Table 2, we compare our EDRNet framework to the Audiovisual Transformer (AVT) (Lin and Wang 2020) and Positive Sample Propagation (PSP) Network (Zhou et al. 2021) on the SEL and WSEL tasks on O-AVE and C-AVE datasets. Our EDRNet framework which leverages TCNs to progressively recognize EPCs, outperforms the PSP on the O-AVE, by **1.06%** and **1.09%** on the WSEL and SEL tasks respectively. On the C-AVE, it outperforms PSP by **1.07%** and **1.11%** on the WSEL and SEL tasks respectively.

Ablation Studies

We report our ablation studies only on O-AVE i.e. the AVE dataset with the *original* test set.

Framework Decoupling: We investigate the contribution of each component and summarize the results for the SEL and WSEL tasks in Table 3. The EDRNet fueled by the SMB video fusion proves to be highly effective, contributing to a **0.83%** increase in SEL accuracy. It bolsters the performance of challenging categories such as "Bus" (+19%) and "Female Speech" (+8%) where conceptual focus is critical to wane off ambient distractions such as traffic and presence of other humans for the former and latter respectively. The success of the EDRNet and the SMB video fusion corroborates the efficacy of EPC-based methods. The LSS loss contributes to a **0.52%** increase for the SEL task and in the *Supplementary Material*, we show that it compactifies the segment features within an event type. B2ILC benefits both tasks, but it influences the WSEL more. For WSEL, the MIL pooling (Eqn. (7)) ensures equal gradient distribution across all segments. Consequently, the difficult segments of commonly confused events (such as "Truck" vs "Bus", "Motorcycle" vs "Racecar", etc.) cannot get prioritized. However, the resolution of easier segments increases the WR of pseudo-positive bags, thereby permitting B2ILC to disambiguate these hard positives (+10.5% and +10% for "Truck" and "Motorcycle" respectively).

EDRNet	B2ILC	$\mathcal{L}_{lss}$	SMBVF	SEL Acc. (%)	WSEL Acc. (%)
✓				76.10	72.98
✓	✓			76.58	73.86
✓	✓	✓		77.10	N/A
✓	✓	✓	✓	77.93	N/A

Table 3: Component-wise ablation study of the EDRNet framework for the SEL and WSEL tasks. SMBVF indicates SMB Video Fusion and B2ILC indicates B2IL Correction.

EDRNet Configurations: We study the influence of the 3 dimensions of EDRNet, i.e., temporal kernel size k, number of (de/re)composition layers L and network width d, on the SEL task. For decompositions, we use a 1D temporal convolution of size k, unit stride and no padding. When applied on a sequence of length N_l at layer $l \in [0, L_{max}]$, the output feature map will be of length $N'_{l+1} = N_l - k + 1$ where $N_0 = N(= 10)$. L_{max} denotes the layer where the net receptive field is maximum, i.e., when $0 < N'_{L_{max}} < k$.

We vary k in the range $[2, 5]$ and correspondingly configure the L as $L_{max}^{k=\{2,3,4,5\}} = \{9, 4, 3, 2\}$ to induce maximum receptive fields (MRFs) for fair comparison. From Fig. 7a, it is evident that $k = 3$ with $L_{max}^{k=3} = 4$ is the optimal kernel size at the MRF. From the perspective of the first layer, the EDRNet makes a localization decision by considering the previous, current and next segment. Next, we fix $k = 3$ and $d = 768$ and vary L in the range $[1, L_{max}^{k=3}]$ where $L_{max}^{k=3} = 4$, to observe how increasing the net receptive field affects performance. From Fig. 7b, we discern that the network performs best at the MRF. Lastly, we freeze $k = 3$ and $L = L_{max}^{k=3} = 4$ and vary d in the range $[256, 1536]$ in steps of 256 to inspect the impact of network width. Fig. 7c shows that increasing d brings benefit until the optimal value ($d = 768$), post which the network starts to overfit.

Impact of SMB Video Fusion: We investigate the effectiveness of the SMB video fusion by quantifying the gains upon varying the extent of augmentation. To produce balanced datasets, we generate S samples per event type where S is varied in the range [50, 400] in steps of 50. The SEL performances by the EDRNet trained on each of the augmented datasets are shown in Fig. 7d. The progressive inclusion of SMB fused videos aids the EDRNet till a certain limit (here, $S = 250$), after which additional videos exposes the network to the same video clips, causing it to overfit.

Modal Branch Isolation Experiments: The clear demarcation of modal branches in the EDRNet permits the examination of branch level performances. Since the EDRNet gated output R_{AV}^{G} is a convex combination of R_A^L and R_V^L, they all share the same feature space. We can extract predictions from the audio (A) and visual (V) branches respectively by replacing R_{AV}^{G} with R_A^L and R_V^L in Eqn. (6). Training various combinations of decomposition (DC) modal branches

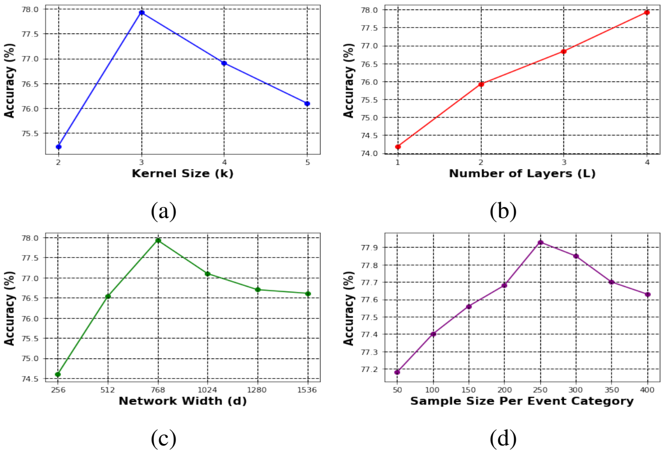


Figure 7: For the SEL task, effect of (a)-(c): Varying the EDRNet's temporal kernel size k, number of (de/re)composition layers L and network width d resp. (d): Varying the extent of SMB video fusion.

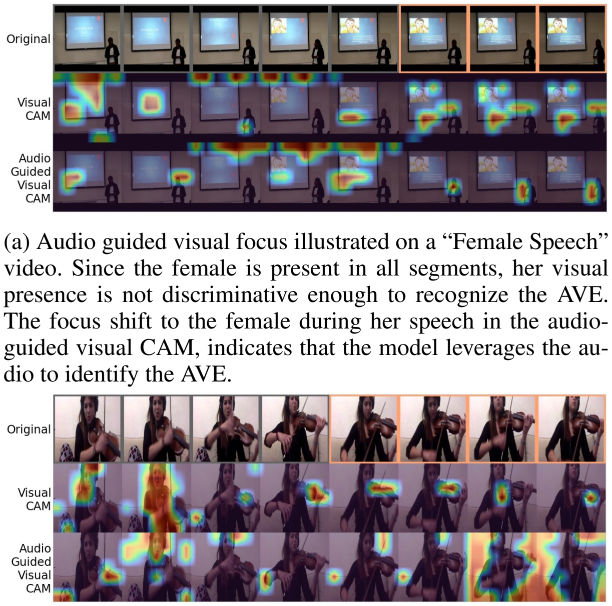
Active DC Branch			Training Configuration	RC Branch Acc. (%)		
A	V	AV		A	V	Gated
✓	-	-	A - Only	65.84	-	-
-	✓	-	V - Only	-	66.93	-
✓	-	✓	A + Dual-Phase Fusion	76.11	-	-
-	✓	✓	V + Dual-Phase Fusion	-	75.26	-
✓	✓	-	A + V with Late Fusion	54.38	62.76	75.88
✓	✓	✓	A + V + Dual-Phase Fusion	73.62	72.74	77.93

Table 4: SEL performance of different modal branch configurations. We train different decomposition (DC) branches and measure performance from the corresponding activated recomposition (RC) branches. A, V and AV denote the audio, visual, and the DC audio-visual branch respectively.

enables us to study the response of the activated recomposition (RC) branches. We visualize our study for better clarity in the *Supplementary Material* and summarize our results on the SEL task in Table 4. By enabling only the A (or V) DC branch (rows 1 & 2), the corresponding RC output delivers the baseline performance for A (or V). The inclusion of the AV branch to DC A & V separately (rows 3 & 4) visibly boosts the RC A & V performances due to the triggering of dual-phase modality fusion. Late fusion achieved by enabling both DC A & V and disabling AV (row 5), allows each modal pathway to learn exclusive cues (reflected by low A & V performances) to cover the counterpart's weaknesses and deliver a high gated accuracy. Finally, by training with all DC branches (row 6), the A & V pathways learn sufficient exclusive cues to furnish the best gated accuracy.

Qualitative Analysis

Our EDRNet framework outperforms prior works which leverage different attention mechanisms to perform and visualize AGVA. This raises an important question: do we need explicit attention mechanisms to achieve cross-modal guidance? To investigate, we use the kernel (sized k) weights corresponding to the maximum activation of D_V^1 and visual portion of D_{AV}^1 (refer Eqn (2)) as coefficients to the visual



(a) Audio guided visual focus illustrated on a "Female Speech" video. Since the female is present in all segments, her visual presence is not discriminative enough to recognize the AVE. The focus shift to the female during her speech in the audio-guided visual CAM, indicates that the model leverages the audio to identify the AVE.

(b) Independent visual focus demonstrated on a "Violin Play" video. The visual CAM captures the model utilizing the visual cue of the contact of the violin bow on the violin to recognize the violin play AVE.

Figure 8: CAM visualizations from the EDRNet. Orange bordered segments indicate the presence of FG event.

feature maps (FMs) corresponding to the GAP output $\hat{f}_t^V$. The k-averaged weighted sum of the FMs yields the Class Activation Maps (CAMs) for the visual and audio-guided visual branch respectively. For $k = 3$, $D_V^1, D_{AV}^1 \in \mathcal{R}^{8 \times d_1}$ and hence 8 CAMs are generated for a 10 second video. Fig. 8 presents CAMs for a "Female Speech" and a "Violin Play" video. In the former, when the woman speaks, the focus distinctly shifts to her in the audio guided visual CAMs. The network capitalized on the audio modality to attend on the visual modality, thereby *implicitly* achieving AGVA. However, the sharp spatial focus on the violin from the visual CAMs in the latter suggests that the network is capable of exploiting only a single modality as well. The above alludes that the EDRNet accomplishes on-demand cross modal guidance without explicit attention mechanisms.

Conclusion

In this paper, we proposed the EDRNet to tackle the SEL and WSEL tasks. Unique to the EDRNet is the ability to localize audio, visual, and audio-visual events simultaneously through the structural assembly of individual and combined viewpoints. Propelled by event atomization, the SMB video fusion can augment datasets with semantically similar but spatiotemporally divergent videos. This exposes the model to diverse content and assists in calibrating its focus on the event source to tackle hard positives. The B2ILC heuristic leverages strong FG neighborhoods to stabilize the localization predictions. Our framework achieves state-of-the-art results on both tasks on the AVE dataset and the effectiveness of each module is validated through extensive experiments.

Acknowledgements: We thank Manasa Bharadwaj and Deepak Sridhar for the time and effort that they have put into carefully reviewing the various drafts of our paper.

References

Adavanne, S.; Politis, A.; Nikunen, J.; and Virtanen, T. 2019. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1): 34–48.

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Alberti, C.; Ling, J.; Collins, M.; and Reitter, D. 2019. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*.

Arandjelovic, R.; and Zisserman, A. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 435–451.

Atrey, P. K.; Hossain, M. A.; El Saddik, A.; and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6): 345–379.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353.

Dietterich, T.; Lathrop, R.; and Lozano-Pérez, T. 2001. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence*, 89: 31–71.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

Gunes, H.; and Piccardi, M. 2005. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, 3437–3443 Vol. 4.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R.; and Wilson, K. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kim, H.; and Bansal, M. 2019. Improving visual question answering by referring to generated paragraph captions. *arXiv preprint arXiv:1906.06216*.

Korbar, B.; Tran, D.; and Torresani, L. 2018. Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 7763–7774. Curran Associates, Inc.

Lai, K.; Yu, F. X.; Chen, M.; and Chang, S. 2014. Video Event Detection by Inferring Temporal Instance Labels. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2251–2258.

Lan, Z.-z.; Bao, L.; Yu, S.-I.; Liu, W.; and Hauptmann, A. G. 2012. Double Fusion for Multimedia Event Detection. In Schoeffmann, K.; Merialdo, B.; Hauptmann, A. G.; Ngo, C.-W.; Andreopoulos, Y.; and Breiteneder, C., eds., *Advances in Multimedia Modeling*, 173–185. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-27355-1.

Lea, C.; Flynn, M.; Vidal, R.; Reiter, A.; and Hager, G. 2017. Temporal Convolutional Networks for Action Segmentation and Detection. 1003–1012.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3889–3898.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Lin, Y.; and Wang, Y. 2020. Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization. In *ACCV*.

Liu, Y.; Ma, L.; Zhang, Y.; Liu, W.; and Chang, S.-F. 2019. Multi-granularity generator for temporal action proposal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3604–3613.

McGurk, H.; and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*, 264(5588): 746–748.

Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H. G.; and Ogata, T. 2015. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4): 722–737.

Oh, S.; Mccloskey, S.; Kim, I.; Vahdat, A.; Cannons, K. J.; Hajimirsadeghi, H.; Mori, G.; Perera, A. G.; Pandey, M.; and Corso, J. J. 2014. Multimedia Event Detection with Multimodal Feature Fusion and Temporal Concept Localization. *Mach. Vision Appl.*, 25(1): 49–69.

Owens, A.; and Efros, A. A. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 631–648.

Parascandolo, G.; Huttunen, H.; and Virtanen, T. 2016. Recurrent neural networks for polyphonic sound event detection in real life recordings. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ramaswamy, J. 2020. What Makes the Sound?: A Dual-Modality Interacting Network for Audio-Visual Event Localization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4372–4376.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*, 234–241. Springer International Publishing.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Schwartz, J.-L.; Berthommier, F.; and Savariaux, C. 2002. Audio-visual scene analysis: evidence for a" very-early" integration process in audio-visual speech perception. In *Seventh International Conference on Spoken Language Processing*.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and So Kweon, I. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4358–4366.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*.

Snoek, C. G.; Worring, M.; and Smeulders, A. W. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 399–402.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 5998–6008. Curran Associates, Inc.

Xu, Z.; Yang, Y.; and Hauptmann, A. G. 2015. A discriminative CNN video representation for event detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1798–1807.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, 570–586.

Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive Sample Propagation along the Audio-Visual Event Line. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.