Dual Perspective Network for Audio-Visual Event Localization

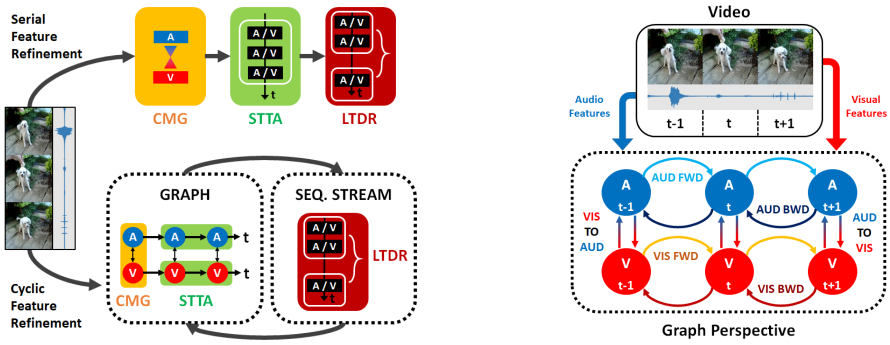
Varshanth Rao[1], Md Ibrahim Khalil[1], Haoda Li[1,2], Peng Dai[1], and Juwei Lu[1]

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

Abstract. The Audio-Visual Event Localization (AVEL) problem involves tackling three core sub-tasks: the creation of efficient audio-visual representations using cross-modal guidance, the formation of short-term temporal feature aggregations, and its accumulation to achieve long-term dependency resolution. These sub-tasks are often performed by tailored modules, where the limited inter-module interaction restricts feature learning to a serialized manner. Past works have traditionally viewed videos as temporally sequenced multi-modal streams. We improve and extend on this view by proposing a novel architecture, the Dual Perspective Network (DPNet), that - (1) additionally operates on an intuitive *graph perspective* of a video to *simultaneously* facilitate cross-modal guidance and short-term temporal aggregation using a Graph Neural Network (GNN), (2) deploys a Temporal Convolutional Network (TCN) to achieve long-term dependency resolution, and (3) encourages interactive feature learning using a *cyclic feature refinement* process that alternates between the GNN and TCN. Further, we introduce the Relational Graph Convolutional Transformer, a novel GNN integrated into the DPNet, to express and attend each segment node's relational representation with its different relational neighborhoods. Lastly, we diversify the input to the DPNet through a new video augmentation technique called Replicate and Link, which outputs semantically identical video blends whose graph representations can be linked to that of the source videos. Experiments reveal that our DPNet framework outperforms prior state-of-the-art methods by large margins for the AVEL task on the public AVE dataset, while extensive ablation studies corroborate the efficacy of each proposed method.

1 Introduction

The rise of various multimedia platforms has resulted in the burgeon of videos across various sectors. The presence of various modalities within a video renders it a rich source of information. Videos stemming from real-life scenes often contain the audio and visual modalities in harmony. In order to understand, recognize and reaffirm events in one modality, processing the other modality can become a necessity. This is particularly true in the case of static sound sources



(a) (b)

Fig. 1: (a) Serial vs cyclic feature refinement involved in the execution of the three core AVEL sub-tasks: Cross-Modal Guidance (CMG), Short-Term Temporal Association (STTA), and Long-Term Dependency Resolution (LTDR). Serial feature refinement limits interaction between sub-task modules to a single pass. Valuable information acquired by the later modules are not conveyed to the earlier modules. Cyclic feature refinement alternates between graph and sequential stream perspectives, enriching the modal features by increasing information exchange between the modules. (b) Visualization of the graph perspective of a sample video. Segment-wise audio and visual features are represented by separate nodes, while edges constitute temporally directed and cross-modal relationships

such as an idling car, where it is difficult to recognize an audio-visual event (AVE) of a static car running without the disambiguation provided by the audio modality. The integrated processing of audio and visual modalities has bolstered methods in various tasks such as sound source localization and separation [11, 1, 17], synthesis of audio from visual data/visual data from audio [5, 4], etc.

AVE Localization (AVEL) engulfs the core sub-tasks of (1) establishment of efficient audio-visual representations of segments through cross-modal guidance, (2) formation of short-term temporal associations to discern patches of event sequences and (3) their accumulation to store event contexts and resolve long-term temporal dependencies. Prior works view videos as sequential modal streams and devise separate modules to tackle these sub-tasks. The sub-network of [20] performs (1), that of [10] perform (1) and (2), while that of [26] perform all three. As illustrated in Fig. 1a, these modules seldom interact with each other apart from the order in which the pipeline is constructed, resulting in the limited serial feature refinement of a segment. Differently, we propose to view videos from an additional graph perspective with modal segment representations as nodes and their interconnections through temporally directed and cross-modal relationships as edges, as shown in Fig. 1b. By using GNNs on the video's graph, a node update encourages a modal segment's features to learn from its temporal neighbors as well as its modal counterparts, hence *simultaneously* addressing sub-tasks (1)

and (2). Although deeper GNNs operating on these video graphs have larger temporal receptive fields, their innate nature induces the oversmoothing effect [8], making them undesirable to perform sub-task (3). Instead, we process the sequential stream perspective of a video using Temporal Convolutional Networks (TCNs) to implement long-term dependency resolution. Further, as shown in Fig. 1a, we alternately process the graph and sequential perspectives, allowing the three sub-tasks to co-refine the features to achieve cyclic feature refinement. We term this procedure as dual perspective processing and the corresponding network as the Dual Perspective Network (DPNet).

Since edges represent different relation types, we can leverage relational GNNs to compose relation-specific node updates. Prior relational GNNs such as Relational Graph Convolutional Network [15] and Relational Graph Attention Networks [3] assume that inter-node relationships are independent. Subsequently, the node updates are derived separately from each relational neighborhood because they are treated as isolated groups. However, in a video, the temporally directed and cross-modal connections between audio and visual segments are semantically related, hence breaking the assumption of relational independence. To induce cross-relational learning, we create a novel GNN called Relational Graph Convolutional Transformer (RGCT) which updates a segment node's relational representation by attending on and learning from its temporal and cross-modal neighborhoods.

An issue with the graph perspective is that segment nodes have limited neighbors of temporally directed and cross-modal nature. Given the segment labels, an effective way to enrich the graph representations is through graph expansion by interconnecting similar videos of the same event type. However, the expansion would be restricted to identical segment sequences, possibly with limited semantic context. Inspired by the CutMix technique [23], we devise a novel video augmentation strategy called Replicate and Link, which preserves the event composition and the semantic context of the original segment sequence. The graphs of the replicas and the originals can then be interlinked to achieve neighborhood expansion for each segment, hence allowing for diversified node updates.

We deploy the DPNet framework to tackle the AVEL problem under the supervised and weakly supervised setting on the public AVE dataset [18]. The contributions of our work are summarized below:

- We propose the Dual Perspective Network (DPNet) to alternately process videos as sequential modal streams, and as graphs. Different from prior works, the DPNet design addresses all the sub-tasks of the AVEL problem while achieving cyclic feature refinement.
- We introduce the Relational Graph Convolutional Transformer (RGCT) to update a node's relational representation by attending across the different relational neighborhoods. RGCTs are used in the DPNet to perform cross-modal guidance and short-term temporal aggregation.
- We design the Replicate and Link video augmentation technique to expand the training set by generating semantically identical replica videos, and enrich a video's graph representation through graph linkage with the replica.

– Experiments show that the DPNet outperforms prior works under the considered settings on the AVE dataset.

2 Related Works

Graphs in Temporal Action Localization (TAL): In TAL, prior works exploit a GNN's ability to perform neighborhood aggregation for refining segment or action proposal features. In [24], the action proposal features are treated as nodes and edges imply a significant temporal overlap or small inter-proposal distance. GNNs perform classification and boundary regression on the node features to achieve TAL. In [13], visual segments form the nodes and the edge weights between all segments are learned with a similarity metric. The learnt inter-segment relation assists in co-localization of similar actions. [22] construct a novel GCNeXt block which splits and operates on snippet nodes using two separate graphs to reflect temporal and semantic connectivity. The graphs are then merged and the updated features are used for performing action localization. Differently, to tackle the AVEL task, our graph constitutes segment nodes and temporal and cross-modal edges. Our GNN, the RGCT, refines segment features by attending across entire relational neighborhoods, rather than across constituent nodes.

Audio-Visual Event Localization (AVEL): The AVEL task entails the identification of temporal regions in a video corresponding to events which are both audible and visible. In [10], a unique Audio-Visual Transformer (AVT) produces short-term spatially attended feature maps corresponding to the sound source into an instance attention module to determine the extent of correlation between the audio and visual components. In [21], an audio-guided spatial-channel attention mechanism is used to refine visual features corresponding to the sound source. The audio features and the attended visual features are processed by blocks of cross-modal scaled dot product attention modules to co-refine modal features before performing segment-wise classification. Recently, [26] introduced a Positive Sample Propagation (PSP) module which calculates and thresholds a similarity matrix between all audio and visual segments. The PSP module then limits the refinement of segment features based on only the positively related connections. Different from prior works, we leverage the graph representation of videos to attend on short-term relationships defined according to temporal and cross-modal directions using GNNs and learn long-term relationships on the stream representation using temporal convolutions.

3 Methodology

3.1 Problem Statement

For the AVEL [18] task, each video sequence is split into N non-overlapping segments. The segment level event label is denoted by $y_t = \{y_t^c | y_t^c \in \{0, 1\}, \sum_{c=0}^{C-1} y_t^c =$

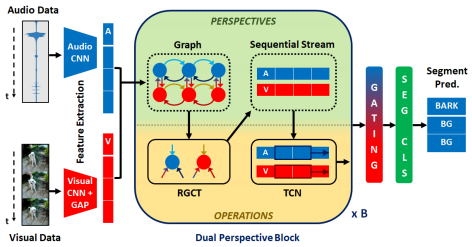


Fig. 2: Illustration of the DPNet used for the AVEL task. Audio and visual features are extracted from a video and fed to a series of Dual Perspective Blocks (DPBlocks). Each DPBlock first processes the graph perspective of the video using an RGCT layer and then processes the sequential stream perspective using a TCN layer. The output audio-visual features are gated and then subject to segment classification

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 [18], we fix the feature extractors and build our architecture on top of these local features. Supervised Event Localization (SEL) and Weakly Supervised Event Localization (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.

3.2 Dual Perspective Network for AVE Localization

We address the three core sub-tasks of AVEL, namely, the establishment of efficient audio-visual representations through cross-modal guidance, the formation of short-term temporal associations, and their accumulation to store event contexts. The first two sub-tasks involve feature interaction within a small temporal neighborhood while the last involves learning to resolve long-term dependencies through the formation of a global understanding. To address the former, we employ GNNs on a video's graph perspective and we tackle the latter using modality-wise temporal convolutions on its sequential stream perspective. A block that sequentially processes both perspectives once is termed as a Dual Perspective Block (DPBlock) and the network with one or more DPBlocks as the Dual Perspective Network (DPNet). We visualize the DPNet architecture for the AVEL task in Fig. 2 and detail its mechanism below.

Graph Perspective: In the graph perspective of a video, a node represents an audio or visual segment's features local to the DPBlock. Concretely, we define

a node representing a segment's features of modality m and time step t within a DPBlock of index b as $n^m_{b,t} = f^m_{b,t}, m \in \{A, V\}$, where $f^m_{b,t}$ represents the input features to the graph perspective layer.

The edges between segment nodes are defined to be temporally directed and cross-modal in nature. Node updates through temporally directed edges enable the encoding of short-term event contexts within the same modality. These contexts can lead to the optimal usage of learnable parameters. E.g., event borders can provide useful cues to neighboring segments regarding how to characterize the start or end of an event. Node updates through cross-modal edges perform cross-modal feature refinement which can assist to achieve a local consensus on an ambiguous AVE. E.g., for static sound sources like idling cars or church bells whose bell movements are occluded, the model can utilize the visual cue of the presence of the static candidate sound source while leveraging the audio signal to confirm the presence of the characteristic sound. All node updates execute simultaneously when the graph is processed using a GNN.

We denote the edge set representing the temporally forward relationships between segments of the same modality m as $\mathcal{E}^m_{b,r_f}$ and temporally backward relationships as $\mathcal{E}^m_{b,r_b}$. Further, we denote the edge set representing the audio to visual and visual to audio relationships between audio and visual segments as $\mathcal{E}_{b,r_{AV}}$ and $\mathcal{E}_{b,r_{VA}}$ respectively. We define $\mathcal{E}^m_{b,r_f}, \mathcal{E}^m_{b,r_b}, \mathcal{E}_{b,r_{AV}}$ and $\mathcal{E}_{b,r_{VA}}$ as:

$$\mathcal{E}^m_{b,r_f} = \{(n^m_{b,t}, n^m_{b,t+1}) | t \in \{1, 2, .., N-1\}\} \tag{1}$$

$$\mathcal{E}^m_{b,r_b} = \{(n^m_{b,t}, n^m_{b,t-1}) | t \in \{2, .., N-1, N\}\} \tag{2}$$

$$\mathcal{E}_{b,r_{AV}} = \{(n^A_{b,t}, n^V_{b,t}) | t \in \{1, 2, .., N\}\} \tag{3}$$

$$\mathcal{E}_{b,r_{VA}} = \{(n^V_{b,t}, n^A_{b,t}) | t \in \{1, 2, .., N\}\} \tag{4}$$

Summarizing the video's graph local to a DPBlock of index b as $\mathcal{G}_b = \{\mathcal{N}_b, \mathcal{E}_b\}$, where $\mathcal{N}_b = \{n^m_{b,t} | m \in \{A, V\}, t \in \{1, 2, .., N\}\}$ represents the modal segment node set and $\mathcal{E}_b = \mathcal{E}^m_{b,r_f} \cup \mathcal{E}^m_{b,r_b} \cup \mathcal{E}_{b,r_{AV}} \cup \mathcal{E}_{b,r_{VA}}$ represents the temporally directed and cross-modal edge set. We process $\mathcal{G}_b$ using a suitable GNN, $\mathcal{F}^{GNN}_b$.

Sequential Stream Perspective: Here, a video is described exclusively by modality-wise temporal sequences. By processing the video within a reference modality, the network learns to assist in forming short-term modality-specific contexts and gradually accumulates these to learn long-term dependencies. This procedure is critical since AVEs can be temporally well spaced. An exemplar case would be the animal sound AVEs which occur discontinuously and in short bursts. The model needs to characterize the entire event and remember the event context to recognize it if it occurs again within the video. We choose Temporal Convolutional Networks (TCNs) over RNNs to process each modal stream, as the former can potentially learn longer sequences than the latter [2].

Formally, we denote the input features of a segment of modality $m \in \{A, V\}$ and time step t to the sequential stream layer of DPBlock index b as $\tilde{f}^m_{b,t}$. The audio and visual streams are denoted respectively as $F^A_b = \{\tilde{f}^A_{b,t}, t \in \{1, 2, .., N\}\}$

and $F_b^V = \{\tilde{f}_{b,t}^V, t \in \{1, 2, .., N\}\}$. We employ TCN layers $\mathcal{F}_{b,A}^{TCN}$ and $\mathcal{F}_{b,V}^{TCN}$ with parameters W_b^A and W_b^V, kernel size k, and the Swish activation [12]:

$$\mathcal{F}_{b,A}^{TCN} = \text{Swish}(\text{TCN}(F_b^A, k; W_b^A)) \tag{5}$$

$$\mathcal{F}_{b,V}^{TCN} = \text{Swish}(\text{TCN}(F_b^V, k; W_b^V)) \tag{6}$$

Since $\mathcal{F}_{b,A}^{TCN}$ and $\mathcal{F}_{b,V}^{TCN}$ operate separately on audio and visual streams, we denote their parallel execution as $\mathcal{F}_b^{TCN}$.

Dual Perspective Network: We first denote the DPBlock of index b as $\mathcal{F}_b^{DPBlock}$ and define it as the sequential execution of $\mathcal{F}_b^{GNN}$ to $\mathcal{F}_b^{TCN}$. The DPNet backbone $\mathcal{F}^{DPNet}$ can be expressed as the sequence of B DPBlocks. Formally,

$$\mathcal{F}_b^{DPBlock} = \mathcal{F}_b^{GNN} \rightarrow \mathcal{F}_b^{TCN} \tag{7}$$

$$\mathcal{F}^{DPNet} = \mathcal{F}_1^{DPBlock} \rightarrow \mathcal{F}_2^{DPBlock} .. \rightarrow \mathcal{F}_B^{DPBlock} \tag{8}$$

For AVEL, we first subject f_t^V to a Global Average Pooling layer, yielding a condensed feature vector $\hat{f}_t^V \in \mathcal{R}^{d_v}$. An FC layer with parameters $W_a \in \mathcal{R}^{d_a \times d_v}$ is applied to f_t^A to yield $\hat{f}_t^A \in \mathcal{R}^{d_v}$. Next, we input $\{\hat{f}_t^A, \hat{f}_t^V\}_{t=1}^N$ to the DPNet backbone. The output audio and visual features of the DPNet backbone are denoted as $\{\hat{\hat{f}}_t^A, \hat{\hat{f}}_t^V\}_{t=1}^N$. We then learn a gating function $\mathcal{F}^G$ through an FC layer parameterized by W_G with a sigmoid activation that operates on a fusion of the features of both modalities. Finally, we apply $\mathcal{F}^G$ to yield the final localization features as a weighted consensus through the convex combination of both the modalities.

$$\mathcal{F}_t^G = \sigma(\text{FC}([\hat{\hat{f}}_t^A, \hat{\hat{f}}_t^V]; W_G)) \tag{9}$$

$$G_t^{AV} = \mathcal{F}_t^G \odot \hat{\hat{f}}_t^A + (1 - \mathcal{F}_t^G) \odot \hat{\hat{f}}_t^V \tag{10}$$

where [.] denotes concatenation and $\odot$ the element-wise product. We transform G_t^{AV} into localization predictions over C classes using an FC layer with a softmax activation.

$$\hat{y}_t = \text{Softmax}(\text{FC}(G_t^{AV}; W_{seg})) \tag{11}$$

The WSEL is formulated as a Multi-Instance Learning (MIL) problem, so we use MIL pooling to aggregate the segment predictions into a video level prediction $\hat{y}$. We use the cross-entropy loss to supervise the SEL and WSEL tasks using the segment-level (y_t) and video-level (y) labels respectively.

3.3 Relational Graph Convolutional Transformer

Earlier, we deduced that the temporally directed and cross-modal edges between audio and visual segment nodes are semantically related, which breaks the assumption of relational independence that fuels the prior GNNs like Relational Graph Convolutional Networks (RGCN) [15] and Relational Graph Attention

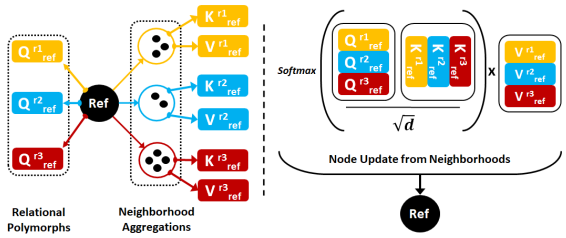


Fig. 3: Visualization of the mechanism of the RGCT. Here, different colors indicate the different relation types. A reference node "Ref" (audio/visual node) is projected into its relational polymorphs as query vectors, while its neigbhorhood aggregations are projected into key and value vectors. The cross-relational scaled dot product attention is used to compose the node update from the relational neighborhoods

Networks (RGAT) [3]. To leverage and extract the semantic relationships between the different edge types, we execute a *cross-relational attention mechanism* via a novel GNN called the Relational Graph Convolutional Transformer (RGCT), as shown in Fig. 3. The RGCT is deployed to the DPNet as the $\mathcal{F}_b^{GNN}$ in a DPBlock.

We simplify the graph notation by omitting reference to DPBlock b as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. We denote the set of indices of the neighbor nodes to a reference node n_i under relation r as η_i^r, where $r \in R$, $R = \{r_{Af}, r_{Ab}, r_{Vf}, r_{Vb}, r_{AV}, r_{VA}\}$ represents the audio and visual temporally directed and cross-modal relationships defined earlier. Next, the neighborhood aggregation of η_i^r is defined as:

$$NA(\eta_i^r) = \frac{1}{|\eta_i^r|} \sum_{j \in \eta_i^r} n_j \tag{12}$$

Nodes can be expressed according to the different relationships it exhibits. These expressions are called *relational polymorphs* and they act as reservoirs of relation-specific details that can be captured from a node's general representation. E.g., a visual node of a person speaking can have the visual temporal forward polymorph encode the presence of the person in the next segment, while the visual to audio polymorph can associate the person's open mouth with the audible speech.

We transform a node n_i into a relational polymorph of type r using an FC layer parameterized by W_Q^r as a query vector $Q_i^r = \text{FC}(n_i; W_Q^r)$. Following the purpose of deriving cross-relational attention, we transform the neighborhood aggregations into key and value vectors parameterized by W_K^r and W_V^r respectively as $K_i^r = \text{FC}(\eta_i^r; W_K^r)$ and $V_i^r = \text{FC}(\eta_i^r; W_V^r)$. We collect the relational polymorphs and neighborhood aggregations of n_i, into single matrix representations as $Q_i = \overset{r}{\|} Q_i^r$, $K_i = \overset{r}{\|} K_i^r$ and $V_i = \overset{r}{\|} V_i^r$, where $\overset{r}{\|}$ is the stack operation over all $r \in R$, $(Q_i^r, K_i^r, V_i^r) \in \mathcal{R}^d$ and $(Q_i, K_i, V_i) \in \mathcal{R}^{|R| \times d}$. We build the cross-relational attention map $Att_i \in \mathcal{R}^{|R| \times |R|}$ and the relation-weighted neigh-

borhood aggregation NA_{att} using scaled dot product attention [19] as follows:

$$Att_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}}) \tag{13}$$

$$NA_{att}(\eta_i) = Att_i V_i \tag{14}$$

To summarize $NA_{att}(\eta_i)$, we average along the relation axis r as $\overline{NA}_{att}(\eta_i) = avg_r(NA_{att}(\eta_i))$. Then, we project n_i to the feature space of $\overline{NA}_{att}$ using an FC layer parameterized by W_1, followed by the Swish activation function. Finally, we update the node n_i to n_i' using an FC parameterized by W_2:

$$\hat{n}_i = \text{Swish}(\text{FC}(n_i; W_1)) \tag{15}$$

$$n_i' = \text{Swish}(\text{FC}(\hat{n}_i + \text{DropOut}(\overline{NA}_{att}(\eta_i)); W_2)) \tag{16}$$

3.4 Replicate and Link Video Augmentation

Eqs. 1-4 reveal that a node in the video's graph possesses a small neighborhood. Small neighborhoods limit node updates and encourage nodes to overfit by creating rigid relation templates. We can alleviate overfitting by increasing the GNN layers to expose nodes to larger neighborhoods. However, on small graphs, this leads to the over-smoothing phenomenon [8]. We tackle this issue by expanding the neighborhood at run-time through the linkage of the graph representations of the original video with that of the semantically identical replicas. We term this video augmentation technique as Replicate and Link and visualize the process through an example in Fig. 4.

Replica Creation: Analogous to action instances in [9], we observe that AVEs can be decomposed into start, continuation, and end sub-events based on temporal progression. Sub-events of the same class type exhibit semantic similarities. E.g., the start sub-event of the helicopter event involves its lift-off from the

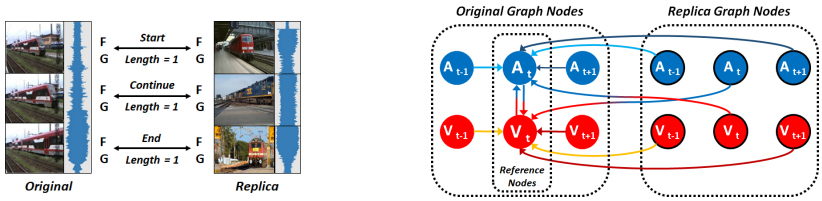


Fig. 4: (a) Illustration of the Replicate operation. A clip of a "Train" AVE is broken down into start, continue and end sub-events. Then, a replica clip of the *same sub-event sequence* is generated. Here, FG indicates the foreground train event. (b) Visualization of the Link operation wherein reference audio and visual nodes receive relevant edge connections from both the original and replica graphs

Algorithm 1: Replica Creation

Input: Original video O of event type e, start, continue and end sub-event databases

Output: Replica video R_O of event type e

Identify a sub-event sequence for O as $SEQ_O = se_1, se_2, ..se_{N_{se}}$, such that there exists sub-event samples of matching segment length for each $se_i \in SEQ_O$, in the se_i database

Initialize R_O to None

for $i \leftarrow 1$ *to* N_{se} **do**

  Choose a random sample v from se_i database matching the segment length of se_i

  Append v to R_O

end

helipad while its end sub-event often involves its landing and the termination of blade rotation. We propose that sub-event segments from different videos of the same event type can be swapped to synthesize semantically identical videos called replicas.

For each training set video of event type e, we first identify and extract the start and end sub-event segments using a one-segment context window around the event border. Next, we identify the continue sub-event segments as those which are wedged between a start and end sub-event. We copy and decompose a continue sub-event of length L into smaller continue sub-events of length $1, 2, .., L-1$. We store the respective sub-events into separate sub-event databases. Then, given an original video and the sub-event databases of event type e, we generate the replica using Algorithm 1. The discontinuity in context introduced by stitching sub-events from different videos allows the network to hone in on the discriminative features specific to the sound source.

Graph Linkage: Given replica videos of identical event sequences, we mutually expand the graph representations of the original and replica videos through graph linkage. Formally, given the graphs of the original and replica as $\mathcal{G}^{orig} = \{\mathcal{N}^{orig}, \mathcal{E}^{orig}\}$ and $\mathcal{G}^{rep} = \{\mathcal{N}^{rep}, \mathcal{E}^{rep}\}$, we merge $\mathcal{G}^{orig}$ and $\mathcal{G}^{rep}$ and then add temporally directed and cross-modal edges between $\mathcal{N}^{orig}$ and $\mathcal{N}^{rep}$ to yield the expanded graph as $\mathcal{G}^{link} = \{\mathcal{N}^{link}, \mathcal{E}^{link}\}$. Here, $\mathcal{N}^{link} = \mathcal{N}^{orig} \cup \mathcal{N}^{rep}$ and $\mathcal{E}^{link}$ is defined below given that $m \in \{A, V\}$:

$$\tilde{\mathcal{E}}^m_{r_f} = \{(\tilde{n}^m_t, \tilde{n}^m_{t+1}) | t \in \{1, 2, .., N-1\}, \tilde{n}^m_t \in \mathcal{N}^{link}\} \tag{17}$$

$$\tilde{\mathcal{E}}^m_{r_b} = \{(\tilde{n}^m_t, \tilde{n}^m_{t-1}) | t \in \{2, .., N-1, N\}, \tilde{n}^m_t \in \mathcal{N}^{link}\} \tag{18}$$

$$\tilde{\mathcal{E}}_{r_{AV}} = \{(\tilde{n}^A_t, \tilde{n}^V_t) | t \in \{1, 2, .., N\}, \tilde{n}^m_t \in \mathcal{N}^{link}\} \tag{19}$$

$$\tilde{\mathcal{E}}_{r_{VA}} = \{(\tilde{n}^V_t, \tilde{n}^A_t) | t \in \{1, 2, .., N\}, \tilde{n}^m_t \in \mathcal{N}^{link}\} \tag{20}$$

$$\mathcal{E}^{link} = \tilde{\mathcal{E}}^m_{r_f} \cup \tilde{\mathcal{E}}^m_{r_b} \cup \tilde{\mathcal{E}}_{r_{AV}} \cup \tilde{\mathcal{E}}_{r_{VA}} \tag{21}$$

Through graph linkage, we diversify the feature space of the aggregated neighborhoods (refer Eq. 12) and correspondingly influence the node update in Eq. 16.

4 Experiments

Dataset and Evaluation Metrics: The AVE dataset [18] is a subset of the AudioSet [6] containing 4143 videos covering 28 real-life event classes such as human speech, vehicle sounds, musical performances etc. Each video is evenly partitioned into 10 segments and each segment is 1 second long. Event labels are available at the segment and video level. AVEs are both audible and visible and spans for at least two seconds. We adopt the same train/validation/test split as [18]. Recently, [26] corrected the annotations for some test videos and report their performance on this corrected test set. We refer to the AVE dataset with the *original* test set as **O-AVE** and the one with the *corrected* version as **C-AVE**. Following all prior works, we evaluate the localization performance using the global classification accuracy of segment 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 [7] and VGG19 [16] networks pretrained on AudioSet [6] and ImageNet [14] respectively. We implement the DPNet using PyTorch Geometric library. The DPNet is built with 4 DPBlocks, each with an RGCT operating first and configured with a dropout probability of 0.2, followed by a TCN layer of kernel size 3. The feature size is set to 768 for all transformations. We train the DPNet using a mini-batch of 48 videos, and use cosine annealing with warm restarts to cycle the learning rates every 20 epochs from 1 to 0.1 till epoch 300, and then to 0.01 till epoch 400. *Only for the SEL setting*, we dynamically generate and link one replica for each video in the mini-batch. For all our experiments, we fix the random seed values for all libraries to ensure reproducible results.

4.1 Quantitative Analysis

Comparisons with SoTA: We compare the AVEL performance of our DPNet on the AVE dataset under the SEL and WSEL settings with Audio-Visual Transformer (AVT) [10], Cross Modal Relation Aware Network (CMRAN) [21] and Positive Sample Propagation (PSP) Network [26]. Unlike prior works, DPNet uses GNNs and TCNs and performs cyclic feature refinement via dual perspective processing. As demonstrated in Table 1, the DPNet with our proposed RGCT outperforms prior works, validating the superiority of cyclic feature refinement. Specifically, DPNet outperforms the previous SoTA, CMRAN, on the O-AVE dataset by **1.53%** on the SEL task and by **1.56%** on the WSEL task. Also, on the C-AVE dataset, it surpasses the previous SoTA, PSP, by **1.88%** on the SEL task and by **1.65%** on the WSEL task.

Effectiveness of the Replicate and Link Augmentation: We investigate the role of the Replicate and Link augmentation technique for the AVEL task under the SEL setting and summarize the ablation in Table 2. We note that the base DPNet performs competitively against prior SoTA methods for the SEL task. We observe that the replication procedure brings a $\sim$**0.6-0.7%** boost. On inspection, we observe large improvements (>10%) on events that have high

Table 1: Performance comparison with SoTAs for the SEL and WSEL tasks on the O-AVE and C-AVE datasets

AVEL Method	Dataset	WSEL Acc (%)	SEL Acc (%)
AVT [10]	O-AVE	70.20	76.80
PSP [26]	O-AVE	72.93	76.84
CMRAN [21]	O-AVE	72.94	77.40
PSP [26]	C-AVE	73.50	77.80
DPNet (Ours w/ RGCT)	**O-AVE**	**74.50**	**78.93**
DPNet (Ours w/ RGCT)	**C-AVE**	**75.15**	**79.68**

Table 2: Ablation study for the Replicate and Link augmentation technique for the SEL task on the O-AVE and C-AVE datasets. Replicate indicates inclusion of the generated replicas during training. Link indicates the interconnection of the graph representations of the original and replica videos

DPNet	Replicate	Link	O-AVE SEL Acc. (%)	C-AVE SEL Acc. (%)
✓			77.50	78.08
✓	✓		78.08	78.78
✓	✓	✓	78.93	79.68

scope to focus on common event contexts rather than the sound source. Example contexts include the green fields where *horses* ride, uniformly colored walls where *clocks* sit, and surrounding traffic with *buses* in between. Further, by applying the link operation, we derive an additional **∼0.9%** increase in overall performance with major improvements visible on AVEs of rodents (+17%), female speech (+8%), and motorcycle (+7%). We observe that these categories benefit from a richer sound source localization due to the feature interpolation achieved during the neighborhood aggregation on the expanded graph representation.

Perspective Combinations: Here, we investigate the influence of each perspective by analyzing the performance of similarly sized networks which cover various perspective combinations. Results are presented in Table 3. Rows 1 and 2 respectively denote the RGAT [3] and RGCT only networks which process only the video's graph perspective. Row 3 denotes a TCN only network operating separately on the audio and visual streams. Rows 4 and 5 present the networks which respectively process the sequential stream (TCN) to graph perspectives (RGCT) and vice versa. The DPNet with Parallel Perspective Block (PPBlock) in row 6 performs parallel processing of both perspectives within a block using the split-transform-merge strategy, instead of the serial style we follow in the DPBlock. Finally, rows 7 and 8 indicate the DPNet with different RGCT-TCN order within a DPBlock as defined in Equation 7. For all networks, the gating mechanism described in Equations 9 and 10 performs the feature fusion for localization. We perform hyperparameter tuning separately on each network to extract the best individual performances.

Within the graph-only perspective setting (rows 1 and 2), we observe that the proposed RGCT only network significantly outperforms the RGAT only network, highlighting the importance of executing the RGCT's cross-relational attention

Table 3: Ablation study on the various networks tailored for different perspective combinations. Performance is reported for the SEL and WSEL tasks on the O-AVE dataset. For the SEL task, Base denotes performance of the network alone, Rep. denotes addition of replica videos during training, and Link denotes the inclusion of the link operation

	Network	Perspective Setting	WSEL	SEL Acc (%)		
			Acc. (%)	Base	+ Rep.	+ Link
1.	RGAT Only	One	51.83	56.11	58.91	60.47
2.	RGCT Only	Perspective	59.22	63.73	64.72	67.13
3.	TCN Only	Only	70.20	74.60	76.10	N/A
4.	TCN → RGCT	Two	59.40	64.17	66.80	70.08
5.	RGCT → TCN	Perspectives	65.52	65.37	70.50	71.07
6.	DPNet w/ PPBlock	Block-wise	71.42	74.82	76.32	77.36
7.	DPNet w/ Graph Second	Two	73.70	76.55	78.01	78.60
8.	**DPNet w/ Graph First**	Perspectives	**74.50**	**77.50**	**78.08**	**78.93**

mechanism for the semantically related relationships. Additionally, we discover that the low results obtained using graph-based (sub)networks (rows 1, 2, 4, and 5) are caused by the oversmoothing effect by the GNN on the localization features which results in similar event predictions for many continuous segments within a video. Similarly, we observe that although the TCN layer in the last DPBlock of the DPNet in row 7 produces discriminative features, the subsequent RGCT layer smooths them across the temporal vicinity before localization, reducing the segment-wise localization performance. Within the DPNet designs (rows 6-8), although the DPNet w/ PPBlock is competitive, it falls short to that of DPNet w/ DPBlock (row 7 and 8). We attribute this to the reduced interaction between the parallel TCN and RGCT layers within a PPBlock as opposed to the richer interaction achieved during their sequential execution within a DPBlock of the DPNet. Finally, we observe that the inclusion of the Replicate and Link augmentation boosts the performance of all methods wherever applicable, with larger increments visible when the base network performance is relatively low.

4.2 Qualitative Analysis

For each relation r of a relevant audio/visual node n_i, we utilize the Class Activation Map [25] algorithm to visualize the Feature Activation Map (FAM) for each relational polymorph, by taking the overall maximum activation (to avoid region inversions) of the query vector Q_i^r of the RGCT in the first DPBlock. Although FAMs here cannot be compared directly to attention maps of prior work, they offer insights into the model's decision-making process. In Fig. 5, we plot the FAMs of the relational polymorphs for segments from videos of different AVE types. We observe that temporally forward polymorphs focus on locating the spatial region(s) corresponding to the actual parts of the sound source, while the temporally backward polymorphs often concentrate on ensuring the presence of the sound source itself. E.g., the audio forward polymorph for the mandolin AVE captures the contact of the player's fingers with the mandolin, and that of

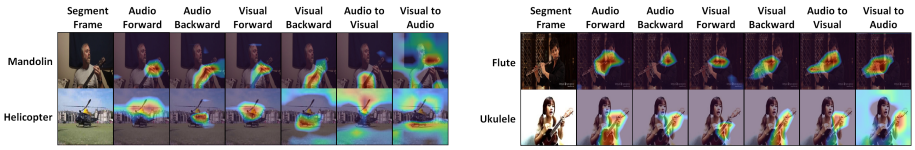


Fig. 5: Visualizations of the Feature Activation Maps from query projections of the audio and visual nodes into relational polymorphs. Each relational polymorph hones its focus onto spatial regions relevant to its semantic functionality, contributing to a rich node update

the helicopter AVE focuses on the rotating blades. In contrast, the audio backward polymorph of the mandolin AVE targets the entire mandolin, and that of the helicopter focuses on the helicopter's body. Similar patterns can be discerned from the visual temporal polymorphs, although less consistently. Additionally, we perceive that the cross-modal polymorphs disseminate information about the sound source from the source modality's perspective. E.g., the audio to visual polymorph for the ukulele AVE focuses on both the player's mouth and ukulele since the person is singing and playing simultaneously. Similarly, for the flute AVE, both the player's mouth and the flute are targeted. In contrast, the visual to audio polymorphs rather focuses on the player's hand contact with the flute and ukulele. It is lucid that the model can focus on different visual regions via the relational polymorphs and this focus is calibrated according to the semantic functionality of the relation type.

5 Conclusion

In this paper, we proposed the DPNet to perform the AVEL task on a video by alternating between its sequential stream and the graph perspectives. By doing so, we achieve cyclic feature refinement between the modules performing cross-modal guidance, short-term temporal aggregation, and long-term dependency resolution. The RGCT was introduced to operate on the graph perspective and achieve cross-relational attention between the relational polymorphs of each node and its relational neighborhoods. The visualizations plotted in the qualitative analysis corroborate that the relational polymorphs implement focus on different spatial regions to propagate relation-specific information during the node update. For the SEL task, the Replicate and Link video augmentation technique enlarged the AVE dataset through the production of semantically identical video replicas and expanded the source video's graph through the interconnection with that of the replica's. Ablation studies demonstrate that both the Replicate and Link operations are effective in assisting the model for the SEL task. Additionally, we validate the superiority of the DPNet structure over other network designs which can operate on both video perspectives. Lastly, comparison results show that the DPNet framework outperforms prior methods in both the SEL and WSEL tasks by large margins.

References

1. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 435–451 (2018)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Busbridge, D., Sherburn, D., Cavallo, P., Hammerla, N.Y.: Relational graph attention networks (2019)
4. Chatterjee, M., Cherian, A.: Sound2sight: Generating visual dynamics from sound and context. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 701–719. Springer International Publishing, Cham (2020)
5. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. In: ECCV (2020)
6. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017)
7. 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., Wilson, K.: Cnn architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017), https://arxiv.org/abs/1609.09430
8. Li, Q., Han, Z., Wu, X.: Deeper insights into graph convolutional networks for semi-supervised learning. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 3538–3545. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16098
9. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
10. Lin, Y., Wang, Y.: Audiovisual transformer with instance attention for audio-visual event localization. In: ACCV (2020)
11. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018)
12. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions (2017)
13. Rashid, M., Kjellström, H., Lee, Y.J.: Action graphs: Weakly-supervised action localization with graph convolution networks. In: Winter Conference on Applications of Computer Vision (2020)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: Gangemi, A., Navigli, R., Vidal, M.E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web. pp. 593–607. Springer International Publishing, Cham (2018)

16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556 (09 2014)
17. Tian, Y., Hu, D., Xu, C.: Cyclic co-learning of sounding object visual grounding and sound separation. In: CVPR (2021)
18. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
20. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2019)
21. Xu, H., Zeng, R., Wu, Q., Tan, M., Gan, C.: Cross-modal relation-aware networks for audio-visual event localization. Proceedings of the 28th ACM International Conference on Multimedia (2020)
22. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
23. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6022–6031 (2019)
24. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: ICCV (2019)
25. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. CVPR (2016)
26. Zhou, J., Zheng, L., Zhong, Y., Hao, S., Wang, M.: Positive sample propagation along the audio-visual event line. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)