2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.01342

Multi-Modal Multi-Action Video Recognition

Zhensheng Shi[1,2] Ju Liang[2] Qianqian Li[2] Haiyong Zheng[2,*] Zhaorui Gu[2] Junyu Dong[1,3] Bing Zheng[2,4]
[1]Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China
[2]Underwater Vision Lab (http://ouc.ai), Ocean University of China
[3]College of Computer Science and Technology, Ocean University of China
[4]Sanya Oceanographic Institution, Ocean University of China

Abstract

Multi-action video recognition is much more challenging due to the requirement to recognize multiple actions co-occurring simultaneously or sequentially. Modeling multi-action relations is beneficial and crucial to understand videos with multiple actions, and actions in a video are usually presented in multiple modalities. In this paper, we propose a novel multi-action relation model for videos, by leveraging both relational graph convolutional networks (GCNs) and video multi-modality. We first build multi-modal GCNs to explore modality-aware multi-action relations, fed by modality-specific action representation as node features, i.e.*, spatiotemporal features learned by 3D convolutional neural network (CNN), audio and textual embeddings queried from respective feature lexicons. We then joint both multi-modal CNN-GCN models and multi-modal feature representations for learning better relational action predictions. Ablation study, multi-action relation visualization, and boosts analysis, all show efficacy of our multi-modal multi-action relation modeling. Also our method achieves state-of-the-art performance on large-scale multi-action M-MiT benchmark. Our code is made publicly available at https://github.com/zhenglab/multi-action-video.*

1. Introduction

Video understanding is a very complex and comprehensive task in computer vision as it aims to recognize activities occurring in a complex environment through complicated hearing and seeing videos [27, 21, 34, 35, 40]. Activities depicted in videos are often made of several actions that may occur simultaneously or sequentially. For example, when the action of "performing" occurs, it is often accompanied by "applauding" and "cheering" actions [35]. *Multi-action video recognition* is such a task that aims to automatically recognize all the actions co-occurring in a video. Although considerable progress has been made in action recognition [44, 46, 51, 6, 47, 52, 62, 12, 31, 58, 11], it's still rather limited on multi-action recognition [35, 41, 60]. In this work, we are going further in more challenging multi-action video recognition to better understand the video.

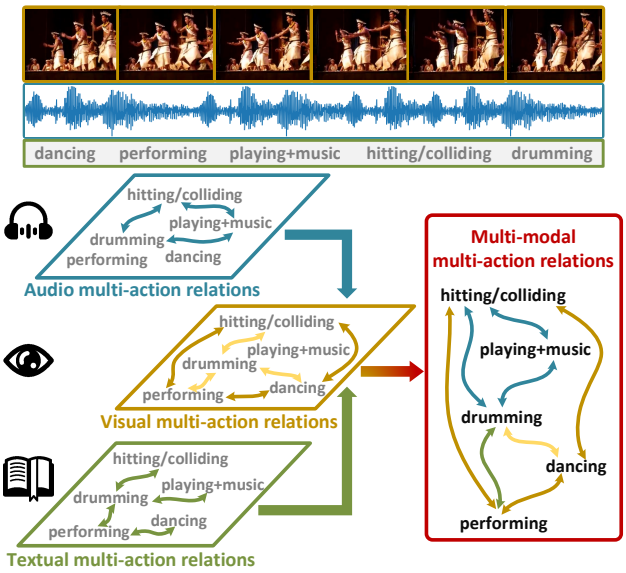

Figure 1: Leveraging multi-modal multi-action relations to recognize all actions in a video (example in M-MiT [35]).

In order to deal with the task of single-action video recognition, more and more efforts are being made to explore the relations between actions and objects from videos [22, 19, 36, 32, 62, 9, 53, 42, 3]. Therefore, to recognize all actions co-occurring in a video for better solving multi-action recognition problem, it would be beneficial and crucial to explore relations among multiple actions, namely, *multi-action relations*. Actually, actions in a video are first presented as visual spatial and temporal frames, also they have strong correlation with synchronous recorded audio, finally they are related to each other in literal meaning (label

*Corresponding author: Haiyong Zheng (zhenghaiyong@ouc.edu.cn).
This work was supported by the National Natural Science Foundation of China under Grant Nos. 61771440, 41927805, and 41776113.

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.01342

text), thus, making full use of these *multi-modal* information in videos (*i.e*., frames, audio, and text) to explore multi-action relations, can contribute greatly to recognize the multiple actions as well as understand the complex videos.

Recent advances in multi-action video recognition mainly focus on, either developing hand-crafted spatiotemporal features [18, 7] (*e.g*., harris corners [23], STIPs [30], optical flow [2], gradient [38]) to train classifiers, or designing 3D convolutional neural network (3D-CNN) architectures to learn discriminative spatiotemporal representation for classification [35, 41]. However, previous works didn't particularly consider the relations among multiple actions in videos. Besides, although multi-modal information has been used to analyze multi-action videos [35], it is only used to extract the feature of corresponding modality (*i.e*., spatiotemporal and auditory features of visual and audio modalities) for fused classification, rather than exploring the multi-modal multi-action relations for more discriminative representation. Thus, how to take full advantage of multi-modal information to better explore multi-action relations is a key point for multi-action video recognition.

Graphs provide a generic way to model real-world relational data [28, 61], and recently, graph convolutional networks (GCNs) [29] have been proved to be very effective at tasks thought to have rich relational structure [64, 39, 56], which might be very helpful for discovering relations among actions from videos. Thus, in this work, we devote to exploring the multi-action relations implied in videos via GCNs by setting multiple actions as graph nodes. Furthermore, due to the unique and important multi-modality property of videos, we build multi-modal GCNs for better modeling modality-specific multi-action relations in videos.

We design our multi-modal GCNs for multi-action video recognition relying on three following observations: (1) visual frames are much more important than other modalities for our daily experience as well as the way we understand the world (more than 80% of information transmitted to the brain is visual) [25], (2) sounds are determined by and informative about the attributes of their actions and we potentially build sound-action mapping from experience inside our brain [15], and (3) our brain can also connect actions with their linguistic labels (meaning words) to create text-action mapping [24, 45]. Figure 1 shows a video example with multiple actions, and visual frames can indicate relations among actions of *performing*, *dancing*, *drumming* and *hitting/colliding* occurring simultaneously or sequentially, also audio from the video will be identified as actions of *playing music*, *drumming* and *hitting/colliding* according to our knowledge of sound-action relations, while the meaning words of occurred actions in this video have their underlying text-action relations semantically, hence, jointing auxiliary textual (underlying) and audio (hearing) relational predictions with primary visual (seeing) relational predictions, can produce more accurate recognition of multiple actions.

In this paper, to address challenging multi-action video recognition, we propose to develop multi-modal GCNs for exploring modality-specific multi-action relations by leveraging graph's powerful relational representation ability and video's rich multi-modal information. Specifically, we construct multi-action graphs with multiple actions as nodes and action co-occurrence probabilities as adjacent matrix, then, we build multi-modal GCNs for exploring modality-aware multi-action relations, fed by modality-specific action representation as node features, *i.e*., spatiotemporal features learned by 3D-CNN, audio and textual embeddings queried from respective feature lexicons, finally, we impose audio and textual relations on spatiotemporal representation to produce respective relational action predictions that are further jointed together with visual relational action predictions to yield final predictions, presenting a novel way of multi-modal joint learning to recognize multiple actions.

Our contributions include: (1) we propose a novel way of taking advantage of relational GCNs and video multi-modality to explore multi-action relations for multi-action video understanding; (2) we devise modality-specific relational GCNs accompanied by multi-modal joint learning for better modeling modality-aware multi-action relations; (3) both ablation study and multi-action relation visualization as well as boosts analysis, show efficacy of our relation modeling, also our method achieves state-of-the-art performance on large-scale multi-action M-MiT benchmark.

2. Related Works

Multi-Action Video Recognition. Multi-action video recognition is such a task that needs to recognize all actions occurring in videos. Early advances in multi-action video recognition mainly focus on developing hand-crafted spatiotemporal features [18, 7] (*e.g*., harris corners [23], STIPs [30], optical flow [2], gradient [38]) to train classifiers. Since the breakthrough of CNNs, the solutions of multi-action video recognition problem are mostly the same as that of action recognition [44, 46, 51, 6, 62, 47, 52, 31, 12, 42, 11, 58], which aim to design effective 3D-CNN architectures to learn discriminative spatiotemporal representation for classification. Wu *et al*. [54] and Wang *et al*. [50] proposed to improve performance of 3D-CNNs in terms of feature and pooling respectively for recognizing multiple actions [43]. Zhang *et al*. [60] dealt with multi-label activity recognition by extracting independent activity-specific features focused on different spatiotemporal regions of a video. Monfort *et al*. [35] annotated a large-scale multi-action M-MiT dataset and concatenated I3D [6] spatiotemporal features with SoundNet [4] auditory features for a single linear layer to rank detected action classes using a new wLSEP loss. Shao *et al*. [41] presented a temporal interlacing network (TIN) to embed the temporal information into the spa-

tial one and learn the information in the two domains once-only. Compared to existing methods for addressing multi-action video recognition, we devote to the solution of exploring the crucial multi-action relations from perspective of natural multi-modality of videos.

Multi-Modal Learning. Multimedia data is often the transmission medium of multiple information, for example, in a video, visual, auditory, and textual information are often disseminated at the same time. Thus, multi-modal learning has gradually developed into the main means of multimedia content analysis and understanding. Among them, visual modality is widely used due to its rich representation ability. In addition, the combination of multiple modalities is commonly considered for strong representation [44, 13, 65, 14, 33, 1, 20]. Different from current multi-modal learning from videos, we provide a novel way of multi-modal joint learning to explore multi-action relations for accurately recognizing all actions in videos according to the observations in real world.

Relation Model. It has already been proved that establishing a relation model is beneficial for understanding videos on action and behavior recognition [22, 19, 36, 32, 62, 9, 42]. Recently, GCN has also been developed to explore relations in videos due to its powerful relation modeling ability. Wang *et al*. [53] used object proposals as nodes to construct a space-time region graph to explore similarity relations and spatial-temporal relations. Wu *et al*. [55] proposed a flexible and efficient Actor Relation Graph (ARG) to capture appearance and position relation between actors for recognizing group activity. Yan *et al*. [57] presented spatial temporal graph convolutional networks (ST-GCN) to exploit spatial relationships among joints for skeleton based action recognition. In our work, rather than only discovering relations from video frames, we set actions as graph nodes to build multi-modal multi-action GCNs for exploring modality-specific multi-action relations in videos.

As we know, recent popular transformer networks are strong relation learners, we also notice their great success in both natural language processing [49, 8] and computer vision [10, 5]. Overall, transformers can also be employed to explore action relations in our framework, that is, we can regard actions as tokens and feed them into transformers for learning multi-action relations, where transformer tokens correspond to GCN graph nodes. The major difference is that transformers miss out on some priori knowledge because they don't have an adjacency matrix whereas GCN does (Section 3.1), but transformers can learn such inherent correlation from more training data, since the intermediate matrix ($softmax(\frac{QK^T}{\sqrt{d_k}})$) from self-attention [49] can be regarded as a dynamically learnable adjacency matrix. Besides, transformers may introduce more parameters, especially for multi-head attention. We will further exploit transformers in our framework in the future.

3. Multi-Modal Multi-Action Relations

3.1. Multi-Modal Multi-Action GCN

Multi-Action GCN. Given a video clip with multiple actions, our goal is to explore the multi-action relations for better recognizing all action categories. GCN has been proved to be very effective at tasks thought to have rich relational structure [64, 39, 56], driving us to dig into GCN for representing multi-action relations. A graph of GCN is made up of nodes which are connected by edges, where things can be represented by nodes while edges can be regarded as connections among them, we thus assign actions as nodes to construct our graphs.

We define a multi-action graph as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ is the set of N nodes representing actions, and $\mathcal{E}$ is the edge set representing co-occurring actions denoted by a binary adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. We formulate correlation dependency of actions as conditional probability $\psi_{ij} = \psi(v_j|v_i)$, which denotes the occurrence probability of action v_j when action v_i occurs. Then we compute ψ_{ij} by counting the occurrence of action pair $\{v_i, v_j\}$ and action v_i from the training dataset, and further we set a threshold t on ψ_{ij} to binarize A_{ij} as initialization, that is, to let $A_{ij} = 1$ if $\psi_{ij} > t$ otherwise $A_{ij} = 0$. By doing so, we actually introduce occurrence probability of actions as adjacent matrix, for constructing multi-action graph in a data-driven way, based on the observation that activities depicted in videos are often made of several actions that may occur simultaneously or sequentially.

We then represent our multi-action GCN using the classic multi-layer fashion with the following layer-wise propagation rule according to [29]:

$$\mathbf{Z}_\zeta^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Z}_\zeta^{(l)}\mathbf{W}_\zeta^{(l)}), \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ is the adjacency matrix of undirected graph $\mathcal{G}$ with added self-connections $\mathbf{I}_N$ that is an identity matrix, $\tilde{\mathbf{D}}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}$ with $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$, $\sigma(\cdot)$ denotes a non-linear activation function (we use Leaky ReLU), ζ represents modality, $\mathbf{W}_\zeta^{(l)}$ is the l^{th}-layer trainable weight matrix, $\mathbf{Z}_\zeta^{(l)}$ is the representation of multi-action relations in the l^{th} layer, and $\mathbf{Z}_\zeta^{(0)} = \mathbf{X}_\zeta$ is the input node features of modality ζ.

So far, we build a general architecture of multi-action GCN with the ability to explore relations among multiple actions. Essentially, multi-action GCN affects each action by aggregating features from its neighbors, thus learns a new representation of an action as relations with other actions. In this way, multi-action relations are gradually aggregated and propagated over multiple layers of GCN counting on the input node features. Actually, multiple actions in videos exist in a multi-modal manner, thus, to better explore multi-action relations, it would be beneficial and

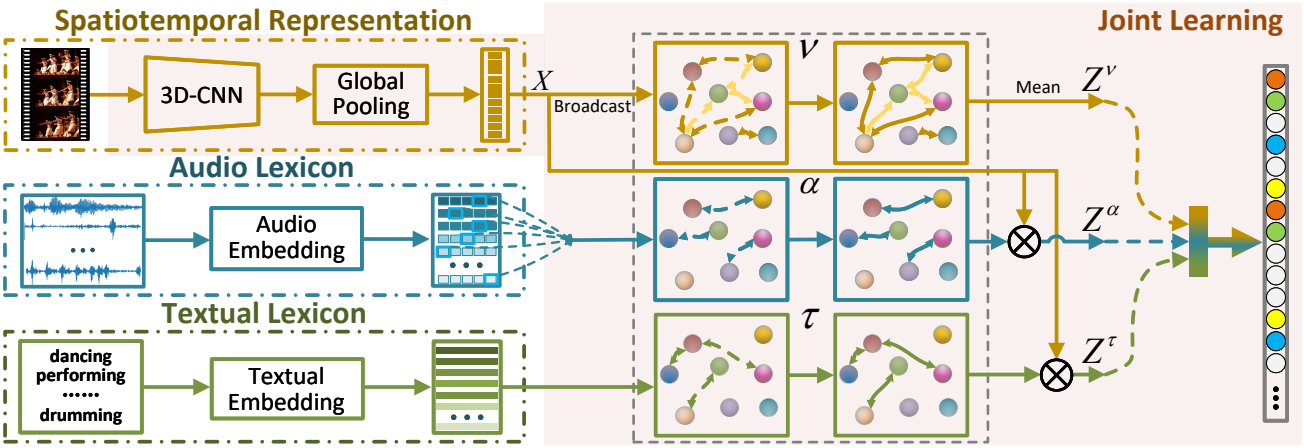

Figure 2: Overview of our method for multi-modal multi-action video recognition. Multi-modal GCNs are designed for exploring modality-aware multi-action relations by feeding modality-specific representation, *i.e.*, spatiotemporal features, audio and textual embeddings, and driven by our multi-modal joint learning for predicting multiple actions more accurately.

crucial to build multi-modal GCNs for leveraging different node features of multiple modalities.

Multi-Modal GCN. Actions in videos are represented in multiple modalities, *i.e.*, visual, audio, and textual, which play different roles in representing the actions. Thus, we construct multi-modal multi-action graphs with three modalities from video dataset, and simply employ a two-layer GCN architecture for each modality in this work ($l = \{0, 1\}$ in Equation 1), where the three modalities are visual ($\zeta = \nu$), audio ($\zeta = \alpha$), and textual ($\zeta = \tau$), respectively.

Spatiotemporal representation of videos contains the most abundant discriminative features for recognizing actions, we therefore employ a 3D-CNN to extract spatiotemporal features and feed them into graph nodes for relation-enhanced classification, resulting in our *visual GCN*. Different from visual modality, audio and text in videos mainly plays an auxiliary role in identifying actions due to their plain representation ability, moreover, corresponding to an action, spatiotemporal features are usually changeful and diverse dynamically, while audio and text are relative stationary. Thus, we design audio-action and text-action feature lexicons for video dataset, and regard them as graph node features for exploring multi-action relations from audio and text modalities to assist visual modality, yielding *audio GCN* and *textual GCN* respectively.

Formally, for *visual* modality, we broadcast the spatiotemporal features $\mathbf{X} \in \mathbb{R}^C$ (C is the dimension) produced by 3D-CNN to $\mathbf{X}_\nu \in \mathbb{R}^{N\times C}$ as node features of N actions, and visual GCN aggregates relation-enhanced features $\mathbf{Z}_\nu^{(2)} \in \mathbb{R}^{N\times N}$, then we take average along the action dimension of $\mathbf{Z}_\nu^{(2)}$ to output visual-modal action prediction $\mathbf{Z}_\nu \in \mathbb{R}^N$. While, for *audio* modality, we denote our lexical audio embeddings as $\mathbf{X}_\alpha \in \mathbb{R}^{N\times P}$ to be graph action features, and the audio-modal multi-action relations $\mathbf{Z}_\alpha^{(2)} \in \mathbb{R}^{N\times C}$ can be propagated from $\mathbf{X}_\alpha$ by audio GCN, finally we impose audio-modal relations $\mathbf{Z}_\alpha^{(2)}$ on spatiotemporal features $\mathbf{X}$ to make audio-modal action prediction $\mathbf{Z}_\alpha = \mathbf{X}(\mathbf{Z}_\alpha^{(2)})^{\mathrm{T}} \in \mathbb{R}^N$. Similarly, for *textual* modality, we represent our lexical text embeddings as $\mathbf{X}_\tau \in \mathbb{R}^{N\times Q}$ for graph actions, so that textual GCN will aggregate text-modal multi-action relations $\mathbf{Z}_\tau^{(2)} \in \mathbb{R}^{N\times C}$ for further text-modal action prediction $\mathbf{Z}_\tau = \mathbf{X}(\mathbf{Z}_\tau^{(2)})^{\mathrm{T}} \in \mathbb{R}^N$.

Until now, we have multi-modal GCNs for exploring modality-aware multi-action relations, fed by modality-specific action features. We next depict our specific way of multi-modal action feature modeling.

3.2. Multi-Modal Action Feature Modeling

Visual-Modal Action Features. Visual modality has strong representations for actions in videos. Recent works on 3D-CNN show powerful performance on parsing and representing visual modality. We hence model visual action features by leveraging 3D-CNN spatiotemporal features.

As we know, in visual modality, actions are dynamically flowing across multiple frames, also they are changeful and diverse. Essentially, by continuous feeding frames, 3D-CNN learns to parse actions via dynamically optimizing spatiotemporal features to be more discriminative, thus finally yielding powerful visual action representation. These visual features, however, implicitly contain relations among multiple actions, which are reasonable and suitable to be action features of visual GCN for further exploring relation-enhanced multi-action representation in visual modality.

Audio and Textual Feature Lexicons. Audio and textual modalities usually act as assistance of visual modality for identifying actions from videos due to their plain rep-

resentation ability. But they still potentially contain audio-action and text-action relations. Thus we exploit audio and textual modalities by modeling their modality-specific action features for audio and textual GCNs respectively, which aggregate modality-specific multi-action relations to further enhance the discriminative spatiotemporal features.

For multi-action video dataset, audios and actions are many-to-many mapping, that is, one audio may correspond to multiple actions and one action may correspond to multiple audios, while, textual labels and actions are one-to-one mapping, namely, one label has the meaning of an action. Therefore, we represent the two modalities by respectively defining many-to-many audio-action and one-to-one text-action feature lexicons for action features of audio GCN and textual GCN. In our work, we employ VGGish [16] and GloVe [37] to represent all audios and label texts of the video dataset as audio and word embeddings for building our audio and textual feature lexicons, respectively.

Formally, we define feature lexicon as a set $\mathcal{L}$ of (f, s) pairs, where a form f is an embedding feature over a finite dimension, and a sense s is the corresponding action from a given set of actions. A feature that corresponds to more than one action is named as *polysemous*, while multiple features that belong to one action are said to be *synonymous*. Then we denote audio and textual feature lexicons as $\mathcal{L}_\alpha$ and $\mathcal{L}_\tau$ respectively, with audio and textual embedding features f_α and f_τ as respective forms while actions s as senses.

The action features of audio and textual GCNs are initialized by querying corresponding lexicons. We model the node features by traversing all senses (actions), and query *synonymous* forms (features) from the lexicon, then the GCNs can reason about "semantic" relations among all modeled actions and features.

3.3. Multi-Modal Joint Learning

We devise multi-modal GCNs to aggregate modality-aware multi-action relations from spatiotemporal feature representation as well as audio and textual feature lexicons, where spatiotemporal features are learned by a 3D-CNN, thus, we propose a joint learning strategy in terms of both model level and representation level involving multiple modalities, namely, *multi-modal joint learning*.

Model Joint Learning. For the whole model learning, we have three modality-specific GCN models ($\mathbf{G}_\nu$, $\mathbf{G}_\alpha$, $\mathbf{G}_\tau$) for relation reasoning and one visual-modal 3D-CNN model $\mathbf{H}$ for spatiotemporal representation learning, where 3D-CNN shares output spatiotemporal features with the three GCNs for aggregating and propagating multi-action relations to produce final action predictions, which will be compared with the real action labels to obtain the model error computed by a loss function, as follows:

$$\mathscr{L}(\mathbf{R}, \mathcal{J}(\mathcal{J}_\nu(\mathbf{H}, \mathbf{G}_\nu), \mathcal{J}_\alpha(\mathbf{H}, \mathbf{G}_\alpha), \mathcal{J}_\tau(\mathbf{H}, \mathbf{G}_\tau))), \quad (2)$$

where $\mathbf{R}$ represents real observations, and $\mathcal{J}$ is a notation denoting the model joint. Subsequently, the modality-specific relational representation will firstly receive gradients of error for updating weights of three GCNs to minimize the loss, and errors will then be propagated from all three GCNs to 3D-CNN via shared spatiotemporal representation for accordingly adjusting its weights. In this way, the whole hybrid model that consists of three GCNs and one 3D-CNN can be trained in a joint learning manner over multiple modalities, such that GCNs are enforced to learn more accurate relational predictions from spatiotemporal representation while 3D-CNN is conducted to model more powerful and relational spatiotemporal features from videos.

Representation Joint Learning. Since every modality has its specific information and representation ability, we take different ways to deal with different modalities. Specifically, dynamic spatiotemporal representation $\mathbf{X}$ is the most influential in recognizing actions from videos thus regarded as the main information flow for model learning, whereas stationary audio-action and text-action lexical representations ($\mathbf{X}_\alpha$ and $\mathbf{X}_\tau$) usually play an auxiliary role in identifying actions thus are considered as the assistant flow. And spatiotemporal representation is gradually learned accompanied by dynamically loading video frames into 3D-CNN, while, audio and textual embeddings queried from corresponding stationary lexicons are simultaneously fed into modality-specific GCNs for assistance. Furthermore, we joint spatiotemporal representation with audio and textual multi-action relations for respective action predictions, and all three modality-specific action predictions are finally fused to produce the final action scores $\mathbf{Z}$, as follows:

$$\mathbf{Z} = \mathbf{G}_\nu(\mathcal{B}(\mathbf{X})) + \mathbf{X}\mathbf{G}_\alpha(\mathbf{X}_\alpha) + \mathbf{X}\mathbf{G}_\tau(\mathbf{X}_\tau), \quad (3)$$

where $\mathcal{B}$ means feature broadcast. By doing so, the information of three modalities is joint to learn better relational representation for recognizing multiple actions.

4. Experiments

4.1. Datasets and Setups

Multi-Moments in Time [35]. We mainly use the recently released Multi-Moments in Time (M-MiT) dataset for experiments, which is considered as a large-scale multi-action dataset for video understanding. M-MiT V1 contains 1.02 million 3 second videos with total 2.01 million labels of 313 action classes annotated from an action vocabulary (*e.g.*, *skateboarding*). In the training set, 553,535 videos are annotated with more than one action, among which 257,491 videos are annotated with three or more actions. M-MiT V2 is the update of V1 with a revision to the action vocabulary, which contains 1 million videos with total 1.92 million labels of 292 action classes, and the training set includes 525,542 videos annotated with more than one action

yet 243,083 videos annotated with three or more actions.

Mini M-MiT. The task of multi-action video recognition is to recognize all actions that occur in videos. However, for M-MiT dataset, we observe that nearly 50% of videos are annotated with only one action label. In order to better explore multi-action video recognition, based on the M-MiT dataset, we intend to build a new dataset which is expected to contain videos annotated with multiple actions for each while retain the integrity of original category. To do this, for the training set, we first remove videos without audio stream, then we randomly select 300 videos for categories with more than 300 videos and choose all the videos of the remaining categories. By doing so, we obtain our "Mini M-MiT" training set with 93,206 videos in 313 action categories. Compared to original M-MiT, our mini M-MiT has only 10% of its data size, such that it's more suitable for quick algorithm development and validation.

IG-65M [17]+Kinetics-400 [27]. IG-65M is a very large-scale pre-training dataset over 65 million public user-generated videos from a social media website, and Kinetics-400 is a classic benchmark for action recognition that contains 246k training and 20k validation videos. In this work, we employ R(2+1)D-34 as our 3D-CNN, pre-trained via finetuning with Kinetics-400 on the released IG-65M pre-trained model (top-1 accuracy: 80.5).

M-MiT Audio and Textual Lexicons. Audio-action lexicon is a set of action-indexed features composed of audio features corresponding to each action of dataset. First, we delete all silent audios in M-MiT to ensure that all audios in the lexicon are valid. Then, we adopt VGGish [16] to extract the features of selected audios with size 3×128. Due to the redundant information in audio data, we further adopt PCA whitening [26] to post-process the extracted features. We finally store the audio features according to action category to obtain our audio-action lexicon.

Similarly, text-action lexicon is a set of action-indexed word features relying on action vocabulary. We use GloVe [37] to extract word embeddings of all actions in the vocabulary of M-MiT, where each action corresponds to one feature vector of size 300, producing our text-action lexicon with word vectors for all actions.

Training and Evaluation. We implement data augmentation and train the model via binary cross-entropy loss optimized by SGD training. Meanwhile, we perform multiple clips testing and use mAP (mean Average Precision), top-1, and top-5 classification accuracy as evaluation metrics. More details are included in *supplementary file*.

4.2. Ablation Study

We conduct ablation studies on our mini M-MiT dataset to validate the efficacy of our multi-modal multi-action relation modeling with pre-trained R(2+1)D-34 as baseline.

Visual GCN vs. Fully-Connected Layer. We start ablation study from our baseline 3D-CNN model R(2+1)D with fully-connected (FC) layer as classifier ($\mathcal{J}(\mathbf{H}, \mathbf{FC})$), which has none of our GCN structures and only involves visual modality. We first replace FC of R(2+1)D with our visual GCN ($\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$) to enhance spatiotemporal features by exploring visual multi-action relations for final action predictions. Table 1 reports the results of jointing different models and involving different modalities, showing that our $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$ model outperforms baseline 3D-CNN model in terms of mAP, top-1, and top-5, so that we can see that our visual GCN does make a positive effect on performance improvement.

model	modality	top-1	top-5	mAP
$\mathcal{J}(\mathbf{H}, \mathbf{FC})$	$\{\nu\}$	52.1	76.0	54.8
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$	$\{\nu\}$	53.3	77.3	55.0
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$	$\{\nu, \alpha\}$	54.3	79.0	58.0
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$	$\{\nu, \tau\}$	54.5	79.7	58.2
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha)$	$\{\nu, \alpha\}$	54.5	79.4	58.2
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\tau)$	$\{\nu, \tau\}$	55.1	79.9	58.5
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha, \mathbf{G}_\tau)$	$\{\nu, \alpha, \tau\}$	55.1	79.8	58.5
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$	$\{\nu, \alpha, \tau\}$	55.0	79.8	58.7

Table 1: Ablation study on multi-modal joint learning.

Multi-Modal Joint Learning. Then, we add an additional modality (audio or textual) on visual modality by jointing 3D-CNN with corresponding GCN (audio GCN or textual GCN), resulting in two joint models $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$ and $\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$ to produce audio and textual action predictions respectively with results reported in Table 1. As it can be observed, by jointing modality-specific GCN with additional modality, both top-1 and top-5 accuracy increase, while mAP is improved significantly with more than 3% boost, indicating the efficacy of our audio and textual GCNs for exploring effective multi-action relations. Besides, we join visual GCN with audio GCN or textual GCN to obtain joint models $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha)$ or $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\tau)$ and fuse the two modality-specific action predictions by removing the absent one from Equation 3, also results in Table 1 show that they leads to additional performance improvement.

Further, we combine all three modalities to yield joint models $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ with absent visual GCN and $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ with all multi-modal GCNs, and Table 1 illustrates that, $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ with three modalities but without visual GCN obtains comparable results (the same top-1 accuracy and mAP) to $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\tau)$ with two modalities yet jointing visual GCN, indicating the effect of visual multi-action relations, while, jointing 3D-CNN with three modality-specific GCNs, for exploring multi-modal multi-action relations, achieves the highest mAP score, demonstrating the efficacy of our multi-modal joint learning. Noting that, our multi-modal GCNs can lead to significant im-

provement by paying a small cost of parameter amount, *e.g.*, our $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$ and $\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$ boosts mAP by 3.2% and 3.4% against baseline 3D-CNN, yet only introducing $0.76M$ and $0.67M$ more parameters. Besides, we try different 3D-CNNs (R3D-18 [47] and I3D-50 [52]) on models ($\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$, $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$, $\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$, $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$), and also yielding effective results (mAP(%)): R3D-18 (45.8, 49.1, 49.5, 50.7) and I3D-50 (53.1, 55.6, 55.8, 57.3).

Moreover, we dig deeper and find out that, for two modalities vs. one modality, our method boosts performance clearly by 3% in mAP, mainly owing to the introduction of additional modality and our design of multi-modal joint learning; while for three modalities vs. two modalities, our method provides slight performance boost, which we consider the reason might be that the plain representation ability of extra auxiliary modality (audio or text) leads to less extra multi-action relation exploration under the same representing mechanism (*i.e.*, GCN and multi-modal joint learning). This inspires us to further improve our method by investigating more modality-specific relational learning and representing fashions (*e.g.*, transformer network) for multi-modal multi-action video recognition.

model	lexicon	top-1	top-5	mAP
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$	1-f	54.3	79.0	58.0
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$	2-f	53.8	79.1	57.8
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$	3-f	53.9	78.8	57.4
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$	*GloVe 300D*	54.5	79.7	58.2
$\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$	*BERT 768D*	54.9	79.5	58.2
$\mathcal{J}(\mathbf{H}, \mathbf{G}_{\alpha,\tau})$	$\{\mathcal{L}_\alpha, \mathcal{L}_\tau\}$	54.9	79.7	58.4

Table 2: Ablation study on audio and textual lexicons.

Audio and Textual Lexicons. We next move on to study single-modal audio or textual feature lexicons.

For audio-action feature lexicon, we traverse all actions to fetch *synonymous* features for each action to initialize node features of audio GCN, so we analyze that how many *synonymous* features to get for an action is better. We thereby conduct ablative experiments by setting the number of *synonymous* features (f) to 1, 2, and 3, and the results shown in Table 2 ($\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$) reveal that, although actions can be represented by many different audios due to their natural many-to-many mapping, it's probably best to select only one audio for representing an action of audio GCN.

For text-action feature lexicon, since actions usually have one-to-one mapping relationship with textual labels (from action vocabulary), we thus study that if different word embedding methods matter. We respectively employ GloVe [37] and BERT [8] to build textual feature lexicons for representing each action with a 300 or 768 dimensional vector. Table 2 ($\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$) shows that, no matter which of GloVe and BERT we use, the accuracy of action predictions is almost the same. Besides, comparing audio joint models $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$ with textual joint models $\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$, the performance is similar, illustrating that the two modalities play similar role in auxiliary recognizing multiple actions.

We also merge audio and textual modalities into one audio-textual modality, by merging audio and textual lexicons to provide audio-textual action representation for one audio-textual GCN, and the results in Table 2 ($\mathcal{J}(\mathbf{H}, \mathbf{G}_{\alpha,\tau})$) demonstrates the superiority of audio-textual modality merging, which actually performs similar to $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ in Table 1. We argue that the merged audio-textual GCN actually tries to explore audio and textual multi-action relations together in one big model, thus achieving similar performance to two separate small audio and textual GCNs.

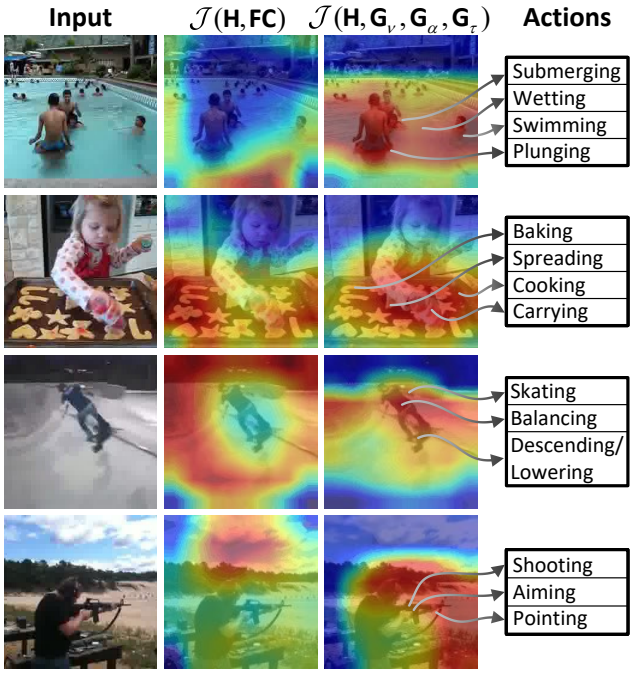

Figure 3: Multi-action Grad-CAM visualization examples with simultaneous actions. The comparison of baseline 3D-CNN model $\mathcal{J}(\mathbf{H}, \mathbf{FC})$ and our multi-modal joint model $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ shows that, thanks to multi-modal joint learning for exploring multi-action relations, our model is able to localize multiple actions present in each scene.

4.3. Multi-Action Relation Visualization

We adopt Gradient-weighted Class Activation Mapping (Grad-CAM) [63] to visualize the learned attention model of 3D-CNN for localizing actions occurring in videos [34, 35], and Figure 3 shows examples with comparison of baseline 3D-CNN model $\mathcal{J}(\mathbf{H}, \mathbf{FC})$ and our multi-modal joint model $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$. As it can be seen, the heatmaps show big difference of the learned 3D-CNN between $\mathcal{J}(\mathbf{H}, \mathbf{FC})$ and $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$, indicating that

our multi-modal joint learning does work for optimizing the 3D-CNN training, also the main difference is that our model is capable of localizing multiple actions presented in each scene. Take the first row for example, $\mathcal{J}(\mathbf{H}, \mathbf{FC})$ is trained to focus on the red region involving *swimming* and *wetting* only, while our model $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ can pay attention to the region including not only *swimming* and *wetting* but also *submerging* and *plunging*, and similar findings can be found in other examples. We argue that, thanks to our model joint learning manner, 3D-CNN benefits a lot from multi-modal GCN models, by receiving backpropagated error from its shared spatiotemporal representation, thus producing more powerful and relational spatiotemporal features for multi-modal GCNs to further better explore modality-specific multi-action relations in videos.

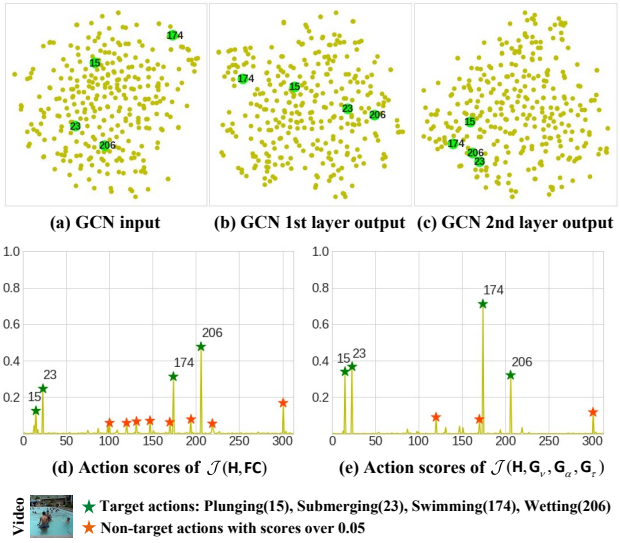

Figure 4: Demonstration of feature changes across GCN layers and action scores for multi-action relations.

We further try to demonstrate our learned multi-action relations. Figures 4(a), (b), and (c) show action embeddings visualization (by t-SNE [48]) indicating feature changes across GCN layers, it can be seen, the target actions (green with number) are gradually aggregated as going through GCN layers, demonstrating the ability to relate multiple actions. Figures 4(d) and (e) show action scores of baseline $\mathcal{J}(\mathbf{H}, \mathbf{FC})$ and our $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$, which illustrate that our model can boost multiple target actions while suppress non-target actions, demonstrating the efficacy of underlying multi-action relation exploration. We also provide visualizations of performance boosts on class-wise (action-wise) AP in Section B (Boosts Analysis) of *supplementary file*.

4.4. Comparison with State-of-the-arts

Table 3 shows comparison with state-of-the-arts on M-MiT datasets, and our model performs best on V1. As V2 was recently released October 2020, no comparison results are available, but we still provide our results for reference.

model-{modality}	back-bone	V1			V2		
		top-1	top-5	mAP	top-1	top-5	mAP
M-MiT-$\{\nu\}$	R50	58.5	81.4	61.7	–	–	–
M-MiT-$\{\nu, \alpha\}$	R50	59.3	82.8	61.8	–	–	–
Ours-$\{\nu\}$	R34	58.6	83.4	61.5	59.5	83.8	62.2
Ours-$\{\nu, \alpha\}$	R34	60.6	85.3	64.0	61.2	85.7	64.4
Ours-$\{\nu, \tau\}$	R34	60.6	85.5	64.1	61.1	85.8	64.5
Ours-$\{\nu, \alpha, \tau\}$	R34	61.2	85.8	64.6	61.7	86.1	65.2

Table 3: Comparison results on M-MiT V1 and V2.

It presents that our best model with three modalities, using a shallower backbone, improves over M-MiT [35] by approximate 3% in mAP. M-MiT adopts a deep SoundNet network for audio feature learning and wLSEP loss with action label statistics, while our visual-audio ($\{\nu, \alpha\}$) model outperform it by 2.2% mAP. Another recent work TIN [41] reports only mAP (62.2) on M-MiT (so we don't list it on the table), which also performs not better than our method. Actually, we can further tap the potential of our solution, via employing more powerful 3D-CNNs or sampling more input frames, *e.g.*, we extend 8-frame to 16-frame for yielding 0.9% boost in mAP on M-MiT V1.

Besides, in this work, we seek to propose a novel way of leveraging multi-modality for multi-action video understanding, and the newly released M-MiT datasets (V1 in 2019 and V2 in 2020) are perfect benchmarks for this study, involving both multi-modality and multi-action as well as their cross references (*e.g.*, *playing music*, *drumming*, and *dancing*). Moreover, we also evaluate our model on Charades dataset [43], which is annotated rarely considering audio multi-action cross reference (MultiTHUMOS [59] ditto), thus we only joint visual and textual modalities and still improve over baseline 3D-CNN model by 2% in mAP. We will discover our model on more datasets in the future.

5. Conclusion

We propose a novel relation model for exploring multi-modal multi-action relations in videos, by leveraging both relational GCNs and video multi-modality. Ablation study, multi-action relation visualization, and boosts analysis, all validate efficacy of our multi-modal multi-action GCNs as well as multi-modal joint learning, on account of the powerful multi-action relation modeling ability. Our method achieves state-of-the-art performance on latest large-scale multi-action M-MiT benchmark. Nevertheless, relations among actions are much more complex, thus more effort is still needed for further digging relying on multi-modal multi-action modeling, and we hope that this work opens up new avenues for multi-action video understanding.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *CVPR*, pages 7558–7567, 2019.

[2] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE TPAMI*, 32(2):288–303, 2008.

[3] Alex Andonian, Camilo Fosco, Mathew Monfort, Allen Lee, Rogerio Feris, Carl Vondrick, and Aude Oliva. We Have So Much in Common: Modeling semantic relational set abstractions in videos. In *ECCV*, pages 18–34, 2020.

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *NIPS*, pages 892–900, 2016.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[7] Johanna Carvajal, Conrad Sanderson, Chris McCool, and Brian C Lovell. Multi-action recognition via stochastic modelling of optical flow and gradients. In *MLSDA*, pages 19–24, 2014.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[9] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-temporal channel correlation networks for action classification. In *ECCV*, pages 284–299, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *CVPR*, pages 203–213, 2020.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016.

[14] Dhiraj Gandhi, Lerrel Pinto, and Abhinav Gupta. Learning to fly by crashing. In *IROS*, pages 3948–3955, 2017.

[15] William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780, 2017.

[17] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, pages 12046–12055, 2019.

[18] Andrew Gilbert, John Illingworth, and Richard Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, pages 925–931, 2009.

[19] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *ICCV*, pages 2470–2478, 2015.

[20] Palash Goyal, Saurabh Sahu, Shalini Ghosh, and Chul Lee. Cross-modal learning for multi-modal video categorization. *arXiv preprint arXiv:2003.03501*, 2020.

[21] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018.

[22] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE TPAMI*, 31(10):1775–1789, 2009.

[23] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, pages 147–151, 1988.

[24] Olaf Hauk, Yury Shtyrov, and Friedemann Pulvermüller. The time course of action and action-word comprehension in the human brain as revealed by neurophysiology. *Journal of Physiology-Paris*, 102(1-3):50–58, 2008.

[25] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019.

[26] Hervé Jégou and Ondřej Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *ECCV*, pages 774–787, 2012.

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[28] Mital Kinderkhedia. Learning representations of graph data: A survey. *arXiv preprint arXiv:1906.02989*, 2019.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, pages 1–14, 2016.

[30] Ivan Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[31] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, pages 7082–7092, 2019.

[32] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, pages 6790–6800, 2018.

[33] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[34] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in Time Dataset: One million videos for event understanding. *IEEE TPAMI*, 42(2):502–508, 2019.

[35] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-Moments in Time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019.

[36] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *CVPR*, pages 1020–1028, 2016.

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[38] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013.

[39] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, pages 593–607, 2018.

[40] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020.

[41] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *AAAI*, pages 11966–11973, 2020.

[42] Zhensheng Shi, Cheng Guan, Liangjie Cao, Qianqian Li, Ju Liang, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. CoTeRe-Net: Discovering collaborative ternary relations in videos. In *ECCV*, pages 379–396, 2020.

[43] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016.

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[45] Rosario Tomasello, Max Garagnani, Thomas Wennekers, and Friedemann Pulvermüller. Brain connections of words, perceptions and actions: A neurobiological model of spatiotemporal semantic activation in the human cortex. *Neuropsychologia*, 98:111–129, 2017.

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[47] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 9(11), 2008.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[50] Jue Wang, Anoop Cherian, Fatih Porikli, and Stephen Gould. Video representation learning using discriminative pooling. In *CVPR*, pages 1149–1158, 2018.

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016.

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[53] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018.

[54] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019.

[55] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, pages 9964–9974, 2019.

[56] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *TNNLS*, pages 1–21, 2020.

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018.

[58] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020.

[59] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 126(2):375–389, 2018.

[60] Yanyi Zhang, Xinyu Li, and Ivan Marsic. Multi-label activity recognition using activity-specific features. *arXiv preprint arXiv:2009.07420*, 2020.

[61] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *TKDE*, 2020.

[62] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018.

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[64] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[65] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, pages 3357–3364, 2017.