Graph-based High-order Relation Modeling for Long-term Action Recognition

Jiaming Zhou[1,5], Kun-Yu Lin[1], Haoxin Li[3], Wei-Shi Zheng[1,2,4*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen 518005, China
[3]School of Electronics and Information Technology, Sun Yat-sen University, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[5]Pazhou Lab, Guangzhou, China
jia_ming_zhou@outlook.com, kunyulin14@outlook.com, lihaoxin05@gmail.com, wszheng@ieee.org

Abstract

Long-term actions involve many important visual concepts, e.g.*, objects, motions, and sub-actions, and there are various relations among these concepts, which we call basic relations. These basic relations will jointly affect each other during the temporal evolution of long-term actions, which forms the high-order relations that are essential for long-term action recognition. In this paper, we propose a Graph-based High-order Relation Modeling (GHRM) module to exploit the high-order relations in the long-term actions for long-term action recognition. In GHRM, each basic relation in the long-term actions will be modeled by a graph, where each node represents a segment in a long video. Moreover, when modeling each basic relation, the information from all the other basic relations will be incorporated by GHRM, and thus the high-order relations in the long-term actions can be well exploited. To better exploit the high-order relations along the time dimension, we design a GHRM-layer consisting of a Temporal-GHRM branch and a Semantic-GHRM branch, which aims to model the local temporal high-order relations and global semantic high-order relations. The experimental results on three long-term action recognition datasets, namely, Breakfast, Charades, and MultiThumos, demonstrate the effectiveness of our model.*

1. Introduction

In the computer vision community, there have been many studies on action recognition. Approaches such as two-stream networks [24], Inflated-3D networks (I3D) [2], and temporal segment networks (TSN) [28] have shown their effectiveness for action recognition. The actions these studies mainly focus on are from pre-trimmed clips of videos and last only a few seconds, which we call short-term actions. However, the videos we are usually exposed to in daily life are more complex long videos, so trimming short-term actions from those long videos is very time-consuming and labor-intensive. More importantly, trimming short-term actions from long videos without considering the possible internal relations between them will prevent action recognition research from achieving the goal of understanding complex human behaviors.

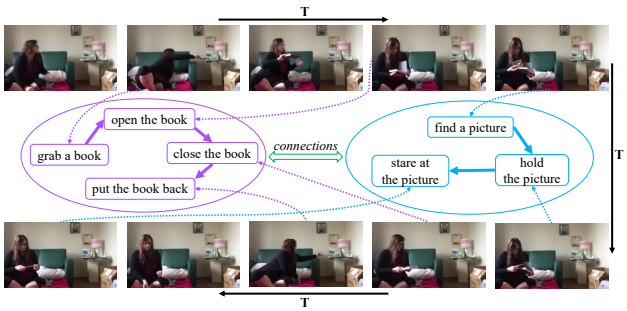


Figure 1. 10 frames selected from the long video *Q948H* in Charades [23] show "*A person sits on a bed, grabs a book off of a table, finds a picture in the book, puts the book back, and stares at the picture*". The purple and cyan cues show two basic relation instances, and there are some connections between them. As the woman opens the book but then closes it quickly without reading it, we can infer that the picture is found from this book. Additionally, the woman does not put the book and picture back together, so she may going to look at the picture. Therefore, the information in these two basic relation instances can affect each other.

In contrast to short-term actions, long-term actions [10, 11, 32] are actions in untrimmed videos that have a very long execution time. A long-term action generally contains multiple sub-actions, among which some complicated relations may exist. The goal of long-term action recognition is to identify the long-term action or to identify all sub-actions that occur during the execution of the long-term ac-

*Corresponding author

tion. Thus, only long-term action category labels or all sub-action category labels need to be provided for long-term action recognition task, which will largely reduce the cost of dataset annotation. Therefore, long-term action recognition provides a feasible way for us to understand complex human behaviors.

In the long-term actions, there are many important visual concepts, *e.g.*, objects, motions, and sub-actions. Various relations may exist among these visual concepts, such as relations between humans and objects, relations between sub-actions, *etc.*, which we call **basic relations**. These basic relations will jointly affect each other during the temporal evolution of long-term actions, which forms the **high-order relations** in the long-term actions. Therefore, exploiting the high-order relations in long-term actions from these basic relations is the key to recognizing long-term actions. In the example from the Charades [23] dataset shown in Figure 1, we can find two visual cues in this long video. The purple cue shows an instance of the basic relation, *i.e.*, "**grab a book** $\rightarrow$ **open the book** $\rightarrow$ **close the book** $\rightarrow$ **put the book back**". The cyan cue shows another instance of the basic relation, *i.e.*, "**find a picture** $\rightarrow$ **hold the picture** $\rightarrow$ **stare at the picture**". There are some connections between these two basic relation instances, which forms the high-order relations. For example, there is a connection between the actions **open the book**, **close the book** in the first basic relation instance and the action **find a picture** in the second basic relation instance, as finding a picture from a book makes it reasonable that a person would open the book but then close it quickly without reading it. Similarly, the actions **close the book**, **put the book back** in the first instance can provide a clue for the action **stare at the picture** in the second instance. Therefore, for long-term action recognition, it is important to model each basic relation well. More importantly, to exploit the high-order relations in the long-term actions, the information from all the other basic relations should be incorporated while modeling each basic relation.

In this work, we propose a novel Graph-based High-order Relation Modeling (GHRM) module, which exploits the high-order relations from the basic relations in the long-term actions for recognition. In GHRM, each basic relation in the long-term actions will be modeled by a graph convolutional module, with each segment in a long video as a graph node. As these basic relations can affect each other and then form the high-order relations during the temporal evolution of long-term actions, GHRM will incorporate the information from all the other basic relations when modeling each basic relation, thus the high-order relations in the long-term actions can be well exploited. To better exploit the high-order relations in the long-term actions along the time dimension, our model constructs the GHRM-layer based on the GHRM module, which considers both the local temporal high-order relations and the global semantic high-order relations by using a Temporal-GHRM branch and a Semantic-GHRM branch, respectively. The Temporal-GHRM branch reasons the graph locally to leverage the temporal context between neighboring segments, while the Semantic-GHRM branch will reason the graph globally to leverage the semantic context between segments without the limitation of temporal distance.

To summarize, the main contributions of this work are:

- A novel GHRM module is proposed to exploit the high-order relations from the basic relations in the long-term actions.
- A GHRM-layer consisting of a Temporal-GHRM branch and a Semantic-GHRM branch is designed to model the local temporal high-order relations and the global semantic high-order relations in the long-term actions.

Our proposed model achieves state-of-the-art performance on three popular long-term action recognition datasets: Breakfast [15], Charades [23], and MultiThumos [31], which demonstrates the effectiveness of our model.

2. Related Work

2.1. Action Recognition

From the perspective of the duration of actions, action recognition can be divided into short-term and long-term action recognition. Short-term actions usually refer to actions that last only a few seconds, while long-term actions are often minute-long.

Short-term Action Recognition. Videos in short-term action datasets (*e.g.*, UCF101 [25], HMDB-51 [16], SomethingV2 [9]) are usually pre-trimmed and last only a few seconds. Early works such as Karpathy *et al.* [13] aimed to recognize short-term actions using 2D CNN. Later, many works [1, 24, 27] utilized the motion features to capture the local dynamics on the time dimension. The temporal cues in action recognition are important: some works [2, 26] extended the 2D CNN to 3D CNN to model the dependencies in the time dimension, and others [5, 6, 19] introduced RNN to learn temporal patterns by treating action recognition as a sequence modeling problem.

Long-term Action Recognition. As more long-term action recognition datasets (*e.g.*, Breakfast [15], Charades [23]) have been proposed, many researchers have turned their attention to long-term action recognition. Sigurdss *et al.* [22] stacked CRF on top of the CNN output to model long-term temporal relations. TRN [34] explored the multi-scale temporal relations among video segments. Non-local networks [29] inserted non-local blocks into 3D CNN to capture long-range dependencies. Timeception [10] built multi-scale convolutional layers on top of CNN features that can recognize complex actions with long-range temporal depen-

dencies. RhyRNN [32] designed a specific RNN structure to ease the gradient flow in a long sequence. However, all these methods struggle to exploit the high-order relations in the long-term actions and ignore either local temporal relations or global semantic relations in the long-term actions.

2.2. Graph Convolutional Networks

The graph convolutional networks (GCN) [14] defined the convolution operation on non-Euclidean structures, which have shown effectiveness in relation modeling and have been widely used in many research areas in computer vision, such as image recognition [4, 17], object detection [18, 21], and action recognition [3, 7, 30]. However, most of these GCN-based approaches are not effective enough for long-term action recognition, as they do not consider the multiple basic relations in the long-term actions. Zhang *et al.* [33] deduced multi-kinds of global semantic relations of an action using multiple graphs, but they did not consider the relations between these graphs while reasoning on each graph. In this paper, our proposed GHRM can incorporate the information from all the other basic relations while modeling each basic relation, thus the information in different graphs will be fused to exploit the high-order relations in the long-term actions.

3. Our Approach

We denote a long video with L frames by $V = \{v_1, v_2, \cdots, v_{L-1}, v_L\}$. Each long video may have two kinds of ground-truths: i) the long-term action category ground-truth is $Y^{la} \in \{0, 1, \cdots, M-2, M-1\}$, where M is the number of long-term action categories in the dataset; ii) the sub-action category ground-truth can be given by the vector $Y^{sa} \in \{0, 1\}^N$, where N is the number of sub-action categories in the dataset. The long-term action recognition task aims to recognize the category of the long-term action Y^{la} or categories of all sub-actions Y^{sa} in a long video.

3.1. Model Overview

In this paper, we propose a model to exploit the high-order relations in the long-term actions. Figure 2 illustrates the framework of our model. For a long video with L frames, our model first uniformly samples T segments where each segment has S consecutive frames. Then we use a feature extractor (*e.g.*, I3D [2]) to extract the segment features $X = \{x_1, x_2, \cdots, x_{T-1}, x_T\}$, where $x_t \in \mathbb{R}^C$ is the feature of the t-th segment and C is the feature dimension. Based on the segment features, our model learns the feature X_{in} by constructing multiple graphs, which will be the input of stacked GHRM-layers to exploit the high-order relations in the long-term actions. In addition, our GHRM-layer contains a Temporal-GHRM branch and a Semantic-GHRM branch to model the local temporal high-order relations and global semantic high-order relations respectively,

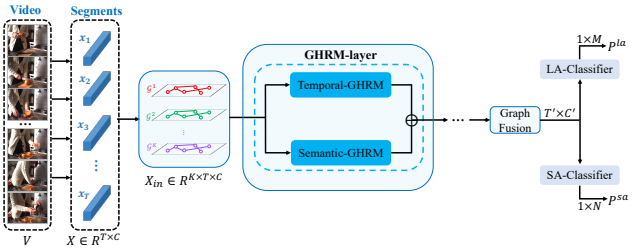


Figure 2. Illustration of our framework for long-term action recognition. We first use a backbone model to extract the segment features from a long video. Then multiple graphs are constructed on the embedded extracted features. Features from all graphs will be fed into GHRM-layer to exploit the high-order relations in long-term actions. Our GHRM-layer contains a Temporal-GHRM branch and a Semantic-GHRM branch to capture the local temporal relations and the global semantic relations respectively. The output features from these two complementary branches will be added together as the input of the next GHRM-layer. After each GHRM-layer, we downsample the segment features in the time dimension to half of the original size. Finally, a graph fusion layer and a classification layer will be applied to recognize the long-term actions.

and features from these two branches will be fused as the input of the next GHRM-layer. Finally, we fuse these graphs and use a classification layer to recognize the long-term actions. In the following, we will elaborate on our model.

3.2. Graph-based High-order Relation Modeling

In this subsection, we describe our Graph-based High-order Relation Modeling (GHRM) module, which aims to exploit the high-order relations in the long-term actions. For each basic relation in the long-term actions, our GHRM constructs a graph to model it. These basic relations will affect each other during the temporal evolution of long-term actions, which forms the high-order relations. GHRM would incorporate the information from all the other basic relations during graph reasoning for each basic relation, thus the high-order relations in the long-term actions can be well exploited. The overall graph structure of our GHRM and how GHRM models the high-order relations will be introduced in the following.

3.2.1 Graph Structure

To model multiple basic relations in the long-term actions, GHRM first constructs a graph on the extracted feature $X \in \mathbb{R}^{T\times C}$ for each basic relation, where each node represents a segment in a long video. We denote these K graphs by $\mathrm{G} = \{\mathcal{G}^1(\mathcal{V}^1, \mathcal{E}^1), \mathcal{G}^2(\mathcal{V}^2, \mathcal{E}^2), \cdots, \mathcal{G}^K(\mathcal{V}^K, \mathcal{E}^K)\}$, where K is the total number of basic relations that we want to model and $\mathcal{G}^i(\mathcal{V}^i, \mathcal{E}^i)$ is the i-th graph used to model the i-th basic relation. In the i-th graph, $\mathcal{V}^i$, and $\mathcal{E}^i$ denote the vertex and edge sets, respectively. Each segment node $x_u \in \mathbb{R}^C$ is

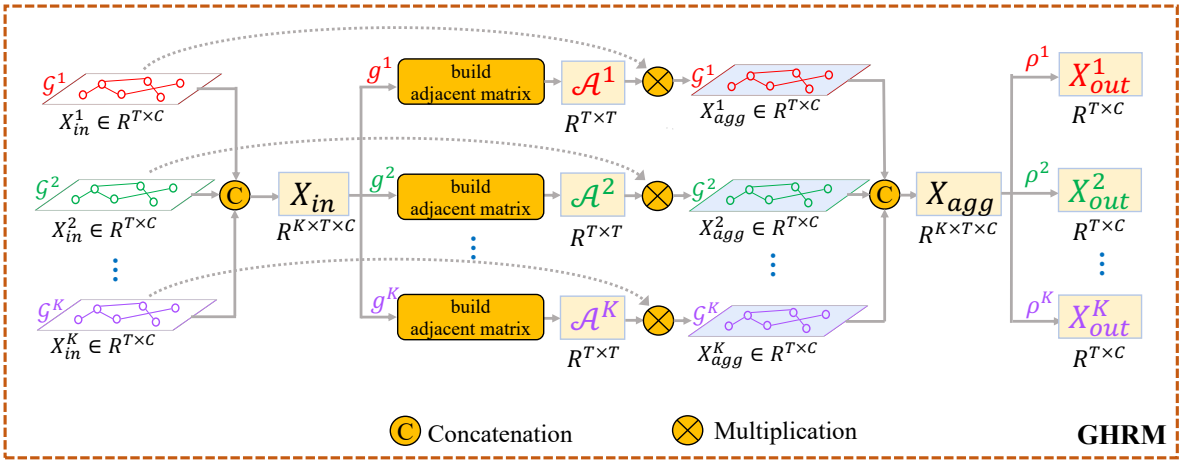


Figure 3. An overview of our proposed GHRM module. There are a total of K graphs for modeling K basic relations in the long-term actions. For each graph $\mathcal{G}^i$, GHRM first calculates its adjacent matrix $\mathcal{A}^i$ with the function g^i which would incorporate the information from all the other graphs. Then, the aggregated feature X^i_{agg} of the segment nodes in each graph $\mathcal{G}^i$ is obtained based on its adjacent matrix $\mathcal{A}^i$. Finally, the new representation of the segment nodes X^i_{out} in each graph $\mathcal{G}^i$ will be calculated by the embedding function ρ^i, which can also incorporate the information from all the other graphs. Best viewed in color.

a vertex in $\mathcal{V}^i$, and $\mathcal{E}^i_{(u,v)} \in \{0,1\}$ in $\mathcal{E}^i$ indicates whether there is an edge between the u-th segment node x^i_u and the v-th segment node x^i_v in the i-th graph.

The graph reasoning process of GHRM is illustrated in Figure 3. For each graph $\mathcal{G}^i(\mathcal{V}^i, \mathcal{E}^i)$, we first calculate its adjacent matrix $\mathcal{A}^i$ by incorporating the information from the other graphs. Then node aggregation will be performed to obtain the aggregated feature X^i_{agg} of the segment nodes in each graph $\mathcal{G}^i$ based on its adjacent matrix $\mathcal{A}^i$. Finally, GHRM utilizes an embedding layer ρ^i, which can also incorporate the information from all the other graphs to obtain the new representation of the segment nodes X^i_{out} in each graph $\mathcal{G}^i$ (the adjacent matrix $\mathcal{A}^i$ and the embedding layer ρ^i are the key to modeling the high-order relations, which will be shown later).

Thus, our GHRM can be formulated as follows for each graph $\mathcal{G}^i(\mathcal{V}^i, \mathcal{E}^i)$:

$$\begin{aligned} X^i_{agg} &= \mathcal{A}^i X^i_{in}, \\ X^i_{out} &= \rho^i(X^1_{agg}, \cdots, X^K_{agg}). \end{aligned} \tag{1}$$

In Equation (1), the input feature $X^i_{in} \in \mathbb{R}^{T\times C}$ represents the segment nodes in the i-th graph, which could be $f^i(X)$ where f^i is a learnable layer used to embed the extracted feature X. $\mathcal{A}^i \in \mathbb{R}^{T\times T}$ is the adjacency matrix of the i-th graph. $X^i_{agg} \in \mathbb{R}^{T\times C}$ is the aggregated node feature in the i-th graph. ρ^i is the embedding layer which would incorporate the information from other graphs to embed the aggregated node feature X^i_{agg} in the i-th graph. $X^i_{out} \in \mathbb{R}^{T\times C}$ is the new hidden representation of the segment nodes in the i-th graph.

3.2.2 High-order Relation Modeling

To exploit the high-order relations in the long-term actions, our proposed GHRM aims to incorporate the information from all the other graphs through the adjacent matrix $\mathcal{A}^i$ and the embedding layer ρ^i in the i-th graph $\mathcal{G}^i$ during graph reasoning for the i-th basic relation.

- Construction of Adjacent Matrix $\mathcal{A}^i$. The adjacent matrix in the graph reflects the degree of correlation between graph nodes, which represents the relation modeled by the graph. Our GHRM uses a graph to model each basic relation in the long-term actions. These basic relations will jointly affect each other and then form the high-order relations during the temporal evolution of long-term actions. Therefore, we consider the influence of the information from all the other graphs when constructing the adjacent matrix $\mathcal{A}^i$ for each graph $\mathcal{G}^i$. For the u-th segment node $x^i_u \in \mathbb{R}^C$ and the v-th segment node $x^i_v \in \mathbb{R}^C$ in the i-th graph, GHRM will calculate the connection strength value $\mathcal{A}^i_{(u,v)}$ between them as:

$$\begin{aligned} \mathcal{A}^i_{(u,v)} &= \begin{cases} \frac{\exp(g^i(x_u)^T g^i(x_v))}{\sum_{w=1}^{T}\exp(g^i(x_u)^T g^i(x_w))}, & \text{if } \mathcal{E}^i_{(u,v)} = 1, \\ 0, & \text{if } \mathcal{E}^i_{(u,v)} = 0, \end{cases} \\ g^i(x_u) &= W^i[(x^1_u)^T, \cdots, (\beta^i(x^i_u))^T, \cdots, (x^K_u)^T]. \end{aligned} \tag{2}$$

When there is an edge between segment nodes x^i_u and x^i_v, *i.e.*, $\mathcal{E}^i_{(u,v)} = 1$ in $\mathcal{E}^i$, $\mathcal{A}^i_{(u,v)}$ is calculated as the inner product with softmax. And when there is no edge between these two nodes, *i.e.*, $\mathcal{E}^i_{(u,v)} = 0$, $\mathcal{A}^i_{(u,v)}$ is set as 0. In Equation (2), $W^i \in \mathbb{R}^{C\times(K\cdot C)}$ is the parameter of embedding layer g^i, which transforms the feature x_u into the i-th relation space, and the feature $x_u \in \mathbb{R}^{K\cdot C}$ is concatenated from

the u-th segment nodes in all graphs (β^i is the embedding layer for segment nodes in the i-th graph), so the connection strength value $\mathcal{A}^i_{(u,v)}$ between segment nodes x^i_u and x^i_v in the i-th graph will be affected by all the other graphs. Therefore, the information from other graphs will be incorporated to exploit the high-order relations.

- Embedding Layer ρ^i. The embedding layer of GCN will embed the aggregated feature of the segment nodes to the new hidden representation. GHRM can incorporate the information from other graphs while embedding the aggregated feature $X^i_{agg} \in \mathbb{R}^{T\times C}$ in each graph $\mathcal{G}^i$ as follows:

$$\begin{aligned}\rho^i(X^1_{agg},\cdots,X^K_{agg}) &= \delta(X_{agg}\mathcal{W}^i)\\ &= \delta([X^1_{agg},\cdots,\gamma^i(X^i_{agg}),\cdots,X^K_{agg}]\mathcal{W}^i),\end{aligned} \tag{3}$$

where δ is a nonlinear function. $\mathcal{W}^i \in \mathbb{R}^{(K\cdot C)\times C}$ is the parameter of embedding layer ρ^i in the i-th graph. $X_{agg} \in \mathbb{R}^{T\times(K\cdot C)}$ is the concatenated aggregated node feature that incorporates information from all the other graphs, thus the high-order relations can be exploited. γ^i is the embedding layer for aggregated node feature X^i_{agg} in the i-th graph.

Remark. By utilizing the adjacent matrix $\mathcal{A}^i$ in Equation (2) and the embedding layer ρ^i in Equation (3) to reason each graph $\mathcal{G}^i(\mathcal{V}^i,\mathcal{E}^i)$, the information from all the other graphs can be incorporated, thus the high-order relations in long-term actions can be well exploited. Compared with GHRM, Vanilla-GCN (a group of standard GCNs [14]) can also use multiple separated graphs to model the basic relations in the long-term actions, which can be seen as a degenerated version of our proposed GHRM. However, Vanilla-GCN will not consider the information exchange between these graphs, which makes itself lack the ability to exploit the high-order relations in the long-term actions.

3.3. Structure of GHRM-layer

To better exploit the high-order relations in the long-term actions along the time dimension, our model constructs the GHRM-layer based on the GHRM module, which takes both the local temporal high-order relations and global semantic high-order relations into consideration. GHRM-layer contains two complementary branches: the Temporal-GHRM branch, which reasons the graph locally to leverage the temporal context between neighboring segments, and the Semantic-GHRM branch, which reasons the graph globally to leverage the semantic context between segments without the limitation of temporal distance.

- Temporal-GHRM. The temporal relations between neighboring segments in videos are important to understand the long-term actions. Temporal-GHRM leverages the temporal context in a long video by defining temporal edges $\mathcal{E}^i$ in each graph $\mathcal{G}^i(\mathcal{V}^i,\mathcal{E}^i)$. For the u-th segment node x^i_u in the i-th graph, Temporal-GHRM restricts its temporal neighboring nodes x^i_v in the i-th graph in a local manner:

$$\mathcal{E}^i_{(u,v)} = \begin{cases}1, & \text{if } |u-v| \le \lfloor \frac{W}{2} \rfloor,\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

where W is the window size that refers to the number of neighboring nodes connected by each segment node.

- Semantic-GHRM. In addition to the temporal relations between neighboring segments, there are also rich semantic relations that may exist in any pair of segments (regardless of the temporal distance between segments) in a long video. Semantic-GHRM leverages the semantic context in a long video using a complete graph:

$$\mathcal{E}^i_{(u,v)} = 1, \tag{5}$$

which means that each pair of segment nodes in each graph has an edge connected.

Our GHRM-layer takes the feature $X^i_{in} \in \mathbb{R}^{T\times C}$ of each graph $\mathcal{G}^i$ together as the input, *i.e.* $X_{in} \in \mathbb{R}^{K\times T\times C}$. The input feature X_{in} will flow to the Temporal-GHRM branch and the Semantic-GHRM branch to capture the local temporal high-order relations and the global semantic high-order relations, respectively. And the output features $X^T_{out} \in \mathbb{R}^{K\times T\times C}$ and $X^S_{out} \in \mathbb{R}^{K\times T\times C}$ from these two branches will be fused into $X_{out} \in \mathbb{R}^{K\times T\times C}$ by adding, which will then serve as the input of the next GHRM-layer.

3.4. Graph Fusion and Classification Header

As GHRM can incorporate the information from other graphs during graph reasoning, the relations modeled by each graph after the stacked GHRM-layers are already high-order relations. Thus, we fuse these graphs in the following effective way. For the segment nodes $X^i \in \mathbb{R}^{T\times C}$ in each graph $\mathcal{G}^i$, we first use a learnable layer $f^i : \mathbb{R}^C \to \mathbb{R}^{\hat{C}}$ to reduce the feature dimension to $\hat{C}$. Then all graphs will be fused into a single graph $\mathcal{G}^F$ with segment nodes $X^F \in \mathbb{R}^{T\times\tilde{C}}$ by concatenating the corresponding node from all graphs, *i.e.*, $X^F = [f^1(X^1),\cdots,f^K(X^K)]$, where $\tilde{C} = K\cdot\hat{C}$. After Graph Fusion, we will feed the segment nodes X^F into the final classification header to predict the category of the long-term action Y^{la} or the categories of all sub-actions Y^{sa} in a long video.

4. Experiments

In this section, we first introduce the metrics used in long-term action recognition and present the details of three long-term action datasets: Breakfast [15], Charades [23], and MultiThumos [31]. Then we will demonstrate the effectiveness of our model and compare it with the state-of-the-arts. After that, we also conduct ablation studies to show the effect of each component of our model. Finally, we compare our GHRM module with Vanilla-GCN quantitatively and qualitatively.

Methods	mAP(%)	Acc(%)
Kinetics [2] pre-trained backbone		
I3D [2]	47.05	58.61
ActionVLAD [8]	60.20	65.48
Timeception [10]	61.82	67.07
VideoGraph [11]	63.14	69.45
Ours	**65.86**	**75.49**
Breakfast fine-tuned backbone		
I3D [2]	61.19	74.83
Ours	**73.89**	**89.01**

Table 1. **Long-term action recognition results on Breakfast.** For the Kinetics pre-trained I3D backbone, our proposed model outperforms the state-of-the-art method VideoGraph [11] by 2.72% in mAP and 6.04% in Acc. Using the fine-tuned I3D backbone, our model achieves 73.89% in mAP and 89.01% in Acc, respectively.

Methods	Modality	mAP (%)
C3D [26]	RGB	10.9
Two-stream [24]	RGB + Flow	18.6
Two-stream + LSTM [24]	RGB + Flow	17.8
ActionVLAD [8]	RGB + iDT	21.0
Temporal Fields [22]	RGB + Flow	22.4
TRN [34]	RGB	25.2
3D ResNet-50 + GCN [30]	RGB + RP	37.5
3D ResNet-101 + NL [29]	RGB	37.5
I3D [2]	RGB	32.9
Timeception [10]	RGB	37.2
VideoGraph [11]	RGB	37.8
Ours	RGB	**38.3**

Table 2. **Long-term action recognition results on Charades.** Our model outperforms the I3D baseline and Non-local Networks by 5.4% and 0.8% in mAP, respectively. Moreover, compared with the state-of-the-art method VideoGraph, it achieves 0.5% improvement in mAP.

4.1. Metrics and Datasets

Recognition metrics. The goal of the long-term action recognition task is to recognize the category of long-term action or the categories of all sub-actions in a long video. Therefore, we use Accuracy (Acc) as the evaluation metric for long-term action classification and Mean Average Precision (mAP) for sub-action classification.

Breakfast. The Breakfast dataset has 1712 breakfast preparation-related long videos. These videos record 52 unique participants, each conducting 10 distinct cooking activities captured in 18 different kitchens. Overall, there are 48 different sub-actions, where each video contains 6 sub-actions and lasts 2.3 minutes on average. We use the same split method as proposed in [10] which has 1357 long videos for training and 355 for testing. As both the long-term action category labels and the sub-action category labels are provided, we evaluate our model with both the Acc and mAP metrics.

Charades. Charades is a large dataset with 9848 annotated long videos, 7985 for training and 1863 for testing. These long videos record 267 people's casual daily activities in 15 types of indoor scenes. On average each video is 30 seconds long and has 6.8 sub-actions. Charades has 157 sub-action classes. As Charades only provides sub-action category labels, we follow Hussein *et al.* [10] to evaluate our model using the mAP metric.

MultiThumos. The MultiThumos dataset is a complex untrimmed dataset which contains 413 long videos, 200 for training and 213 for testing, with a total duration of 30 hours. It is extended from Thumos14 [12] by providing dense and multi-label annotations for the videos. There are a total of 65 sub-actions in MultiThumos, with an average of 11 sub-actions in each video. The metric mAP is used for evaluation on this dataset.

4.2. Implementation Details

We use PyTorch [20] to implement our model. Following the experimental settings in [10], for a long video with L frames, we uniformly sample 64 segments with 8 consecutive frames in each segment, thus a total of 512 frames will be sampled from each long video. The segment features are extracted from the output of the last pooling layer of the I3D network [2], so the size of each segment feature is 1024 with the spatial dimension maxpooled. To make our GHRM module more efficient and lightweight, we adopt a channel sharing strategy on layer g^i in Equation (2) and layer ρ^i in Equation (3) during graph reasoning for each graph $\mathcal{G}^i$. The window size W in Temporal-GHRM is 7. We use stochastic gradient descent with a momentum of 0.9 to optimize our model. The batch size is 32 and the learning rate is 0.5. The weight decay is set as 10^{-4}. Similar to [10], the backbone model in our framework will not be end-to-end trained. See more details in the **supplementary material**.

4.3. Comparison with the State-of-the-art

Breakfast: Table 1 compares the long-term action recognition results of our model and several related methods on the Breakfast dataset by stacking four GHRM-layers with 32 graphs on I3D backbone. As shown in Table 1, for the Kinetics [2] pre-trained I3D backbone, our model outperforms the I3D baseline by 18.81% in mAP and 16.88% in Acc, and outperforms the state-of-the-art method VideoGraph [11] by 2.72% in mAP and 6.04% in Acc. We also fine-tune the I3D on the Breakfast dataset to extract better visual features. Using the fine-tuned I3D as the backbone, our model achieves 73.89% in mAP and 89.01% in Acc, which is a large improvement over the fine-tuned I3D.

Methods	mAP (%)
I3D [2]	72.43
Timeception [10]	74.79
Ours	**79.89**

Table 3. **Long-term action recognition results on MultiThumos.** Our model outperforms the I3D baseline and the state-of-the-art method Timeception by 7.46% and 5.10% in mAP respectively.

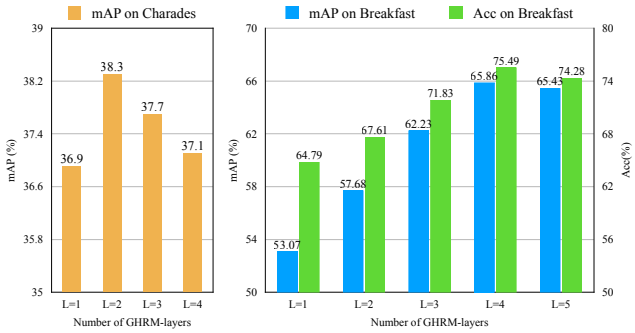


Figure 4. Comparison of the effects of using different numbers of GHRM-layers. Our model achieves the best results on Charades and Breakfast using two and four GHRM-layers respectively.

Charades: We compare our model with both short-term and long-term action recognition methods by stacking two GHRM-layers with 64 graphs on the I3D backbone. As shown in Table 2, our model outperforms the I3D baseline and Non-local Networks [29] which uses a stronger backbone by 5.4% and 0.8% in mAP, respectively. Moreover, our model outperforms the state-of-the-art long-term action recognition method VideoGraph [11], achieving state-of-the-art performance on this large dataset.

MultiThumos: We use MultiThumos as the third long-term action recognition dataset to demonstrate the effectiveness of our model. By stacking four GHRM-layers with 64 graphs on the Kinetics pre-trained I3D backbone, we achieve state-of-the-art performance on MultiThumos. As shown in Table 3, our model can reach 79.89% in mAP, which exceeds the state-of-the-art method Timeception [10] by 5.10% in mAP.

4.4. Ablation Studies

Number of GHRM-layers. Stacking multiple GHRM-layers is effective for modeling the long-term relations in long videos. In Figure 4, we show the effects of using different numbers of GHRM-layers on the Charades and Breakfast datasets. The results show that stacking two and four GHRM-layers are sufficient for the Charades and Breakfast datasets respectively. By stacking more GHRM-layers, the performance on Breakfast can be continuously boosted, while it does not on Charades. The reason is that the average duration of the videos in Breakfast is 2.3 minutes, which is much longer than 0.5 minutes in Charades.

Number of Graphs	mAP on Breakfast	mAP on Charades
K=8	61.2	36.5
K=16	63.1	37.5
K=32	**65.9**	38.0
K=64	65.7	**38.3**
K=128	63.8	38.1

Table 4. mAP(%) comparisons of different numbers of graphs used in GHRM. We achieve the best results on Breakfast and Charades when using 32 and 64 graphs respectively.

Models	mAP on Breakfast	mAP on Charades
I3D	47.1	32.9
Semantic-GHRM	56.4	36.7
Temporal-GHRM	62.9	37.3
Ours	**65.9**	**38.3**

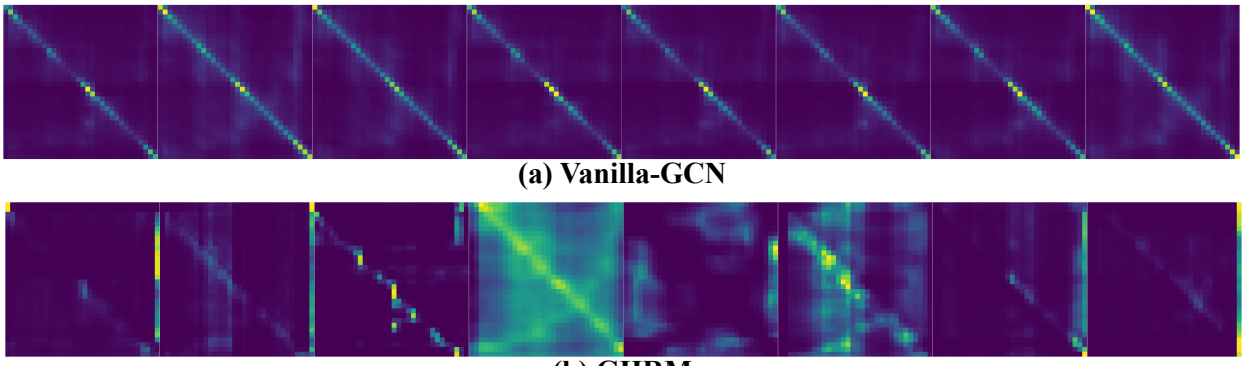
Table 5. The effects of the Temporal-GHRM branch and Semantic-GHRM branch of our model. The mAP(%) results on both Breakfast and Charades drop remarkably when either branch is removed.

Number of Graphs. Our GHRM uses a graph to model each basic relation in the long-term actions, so we analyze the effects of using different numbers of graphs in GHRM on both Breakfast and Charades. As shown in Table 4, our model achieves the best result on Breakfast with 32 graphs, but 64 graphs are needed for achieving the best result on Charades. As the long-term actions in Charades are more diverse, which requires our model to use more graphs on this dataset.

Model Components. Our GHRM-layer contains a Temporal-GHRM branch and a Semantic-GHRM branch that aim to model the local temporal high-order relations and global semantic high-order relations, respectively. We ablate these two branches by removing one of them. As shown in Table 5, the results on both Breakfast and Charades drop remarkably when either branch is removed, which demonstrates that our Temporal-GHRM branch and Semantic-GHRM branch are complementary.

4.5. Comparison with Vanilla-GCN

Both Vanilla-GCN and the proposed GHRM module use multiple graphs to model the basic relations in long-term actions. Our GHRM can exploit the high-order relations in the long-term actions by incorporating the information from all the other graphs when reasoning on each graph, while the Vanilla-GCN fails to exploit the high-order relations due to the lack of information exchange between multiple graphs. To further demonstrate the effectiveness of modeling high-order relation on long-term action recognition, we replace our GHRM module with Vanilla-GCN for quantitative and



(a) Vanilla-GCN

(b) GHRM

Figure 5. Visualizations of the adjacent matrices in Vanilla-GCN and GHRM. Figure (a) shows 8 adjacent matrices in Vanilla-GCN that have very similar patterns. Figure (b) shows 8 adjacent matrices in GHRM that have very diverse patterns.

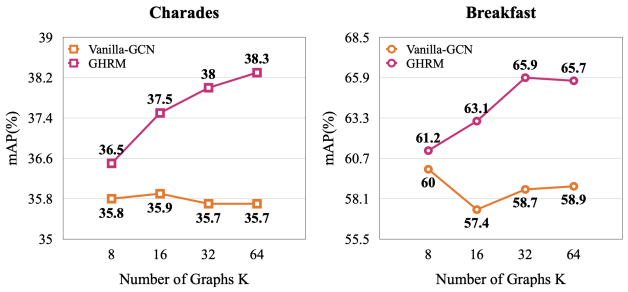


Figure 6. Comparison between GHRM and Vanilla-GCN using different numbers of graphs on both Charades and Breakfast. Regardless of how many graphs are used, GHRM is much better than Vanilla-GCN. And with more graphs used, the performance of our GHRM generally increases, while the performance of Vanilla-GCN does not.

qualitative comparisons.

Quantitative Analysis. In Figure 6, we compare the performance of our GHRM and Vanilla-GCN on both Charades and Breakfast with different numbers of graphs K. As shown in this figure, regardless of how many graphs are used to model the basic relations in long-term actions, our GHRM performs much better than Vanilla-GCN. Moreover, with more graphs used, the performance of our GHRM generally increases, while the performance of Vanilla-GCN does not. These results demonstrate high-order relation modeling is effective to recognize the long-term actions.

Qualitative Analysis. Figure 5 shows 8 adjacent matrices of a video sample produced by Vanilla-GCN and GHRM respectively (randomly selected from the 16 semantic adjacent matrices in both modules). Figure 5 (a) shows 8 adjacent matrices in Vanilla-GCN which have very similar patterns, while Figure 5 (b) shows 8 adjacent matrices in GHRM that have diverse patterns. This phenomenon indicates that different graphs can interact with each other in GHRM, such that different basic relations will be figured out and the high-order relations in the long-term actions can be naturally exploited. See more visualizations in the **supplementary material**.

5. Conclusions

In this paper, we propose a Graph-based High-order Relation Modeling (GHRM) module to tackle the task of long-term action recognition. Long-term actions involve many visual concepts, and there are many basic relations among these concepts. As these basic relations will affect each other and then form the high-order relations during the temporal evolution of long-term actions, our proposed GHRM incorporates information from all the other basic relations when modeling each basic relation, thus the high-order relations in the long-term actions can be well exploited from these basic relations. Furthermore, to better model the high-order relations in the long-term actions along the time dimension, a GHRM-layer consisting of a Temporal-GHRM branch and a Semantic-GHRM branch is designed to model the local temporal high-order relations and the global semantic high-order relations in the long-term actions. The experimental results on three popular long-term action recognition datasets (*i.e.*, Breakfast, Charades, and MultiThumos) have demonstrated the effectiveness of our proposed model.

6. Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(U1911401,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (No. 2020B1515120085, No. 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004).

References

[1] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 6, 7

[3] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[5] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[6] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 2017. 2

[7] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019. 3

[8] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan C. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision*, 2017. 2

[10] Noureldien Hussein, Efstratios Gavves, and Arnold W. M. Smeulders. Timeception for complex action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 7

[11] Noureldien Hussein, Efstratios Gavves, and Arnold W. M. Smeulders. Videograph: Recognizing minutes-long human activities in videos. *CoRR*, 2019. 1, 6, 7

[12] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 2017. 6

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *the 5th International Conference on Learning Representations*. OpenReview.net, 2017. 3, 5

[15] Hilde Kuehne, Ali Bilgin Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 5

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011. 2

[17] Qing Li, Xiaojiang Peng, Yu Qiao, and Qiang Peng. Learning label correlations for multi-label image recognition with graph networks. *Pattern Recognition Letters*, 2020. 3

[18] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for rgb-d salient object detection. In *European Conference on Computer Vision*, 2020. 3

[19] Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, and Cewu Lu. Deep RNN framework for visual sequential applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[21] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[22] Gunnar A. Sigurdsson, Santosh Kumar Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6

[23] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016. 1, 2, 5

[24] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *the 27th International Conference on Neural Information Processing Systems*, 2014. 1, 2, 6

[25] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 2012. 2

[26] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, 2015. 2, 6

[27] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[29] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6, 7

[30] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *European Conference on Computer Vision*, 2018. 3, 6

[31] 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. *International Journal of Computer Vision*, 2018. 2, 5

[32] Tianshu Yu, Yikang Li, and Baoxin Li. Rhyrnn: Rhythmic RNN for recognizing events in long and complex videos. In *European Conference on Computer Vision*, 2020. 1, 3

[33] Jingran Zhang, Fumin Shen, Xing Xu, and Heng Tao Shen. Temporal reasoning graph for activity recognition. *IEEE Transactions on Image Processing*, 2020. 3

[34] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, 2018. 2, 6