

# Anticipating Future Relations via Graph Growing for Action Prediction

Xinxiao Wu,<sup>1</sup> Jianwei Zhao,<sup>2</sup> Ruiqi Wang<sup>1</sup>

<sup>1</sup>Beijing Laboratory of Intelligent Information Technology

School of Computer Science, Beijing Institute of Technology, Beijing, China

<sup>2</sup>School of Information Science and Technology, Northeast Normal University, Changchun, China

{wuxinxiao,wang\_ruiqi}@bit.edu.cn, zhaojw374@nenu.edu.cn

## Abstract

Predicting actions from partially observed videos is challenging as the partial videos containing incomplete action executions have insufficient discriminative information for classification. Recent progress has been made through enriching the features of the observed video part or generating the features for the unobserved video part, but without explicitly modeling the fine-grained evolution of visual object relations over both space and time. In this paper, we investigate how the interaction and correlation between visual objects evolve and propose a graph growing method to anticipate future object relations from limited video observations for reliable action prediction. There are two tasks in our method. First, we work with spatial-temporal graph neural networks to reason object relations in the observed video part. Then, we synthesize the spatial-temporal relation representation for the unobserved video part via graph node generation and aggregation. These two tasks are jointly learned to enable the anticipated future relation representation informative to action prediction. Experimental results on two action video datasets demonstrate the effectiveness of our method.

## Introduction

Action prediction refers to inferring action category labels from partially observed videos that contain incomplete action executions. It is very challenging since it is difficult to exploit sufficiently discriminative information from partial videos to make accurate prediction. Many existing studies (Cai et al. 2019; Wang et al. 2019) learn an enriched feature representation from the partial video by transferring discriminative information from the full video. Several other methods (Zhao and Wildes 2019; Gammulle et al. 2019) generate a feature representation of the unobserved video part to enhance the complete action representation for classification. These methods have achieved promising performance on action prediction, however, such considerable progress has been made without exploring potentially valuable structure information within the video such as the interaction and correlation between different object entities. It is a fact that explicitly modeling the visual object relations in videos has been demonstrated to play a pivotal role in action analysis (Xiao et al. 2019; Zhou et al. 2018; Wang, Li, and Van Gool 2018; Tsai et al. 2019).

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this paper, we investigate how the fine-grained visual relations between objects evolve over both spatial and temporal domains for action prediction. Inspired by the Gestalt law in psychology that human beings have the innate tendency to perceive the incomplete as complete and then they unconsciously attempt to fill in the gap (Farahzad 1998; Ehrenstein, Spillmann, and Sarris 2003; Spillmann 2006), we devote to anticipating the visual relations for the unobserved video part to enable a complete perception of the video, thus facilitating action classification. With this in mind, we propose a graph growing method that involves two tasks: (1) reasoning relations in the observed video part via graph neural networks and (2) synthesizing relations for the unobserved video part via graph node generation and aggregation.

In the relation reasoning task, we propose a spatial-temporal relation reasoning model to extract the spatial relations between objects in still frames and explore how these spatial relations dynamically change over time. Specifically, we use gated graph neural network to perform the spatial relation reasoning within video frames. Each video frame is formulated as a spatial graph where the node denotes an object and the directed edge denotes the spatial relation between two objects. For the temporal relation reasoning, we propose a long short-term graph convolutional network (LST-GCN) to model both the short-term and long-term temporal evolutions of the spatial relation with multi-scale receptive fields. A spatial-temporal graph is built on each video by formulating the spatial graph of each frame as a super node and the temporal relations between frames as directed edges. Consequently, the spatial-temporal relation reasoning is implemented by message propagation on the spatial-temporal graph to learn the evolution of the structural information over both space and time.

In the relation synthesizing task, we propose a relation synthesizing model to make the spatial-temporal graph built in the reasoning task grow into representing the relations of the complete video. The new grown part can be regarded as a *synthesized sub-graph* to represent the relations of the unobserved video frames. Then the built spatial-temporal graph from the observed video part can be called *observed sub-graph*. The observed sub-graph is grown in multiple temporal scales to maintain the varying dynamics of relations with different granularities. For the  $z$ -th scale,  $z$  nodes in

the observed sub-graph are sampled with equal interval and then fed into a graph auto-encoder to generate  $z$  unobserved nodes in the synthesized sub-graph. The representations of the forecasted  $z$  nodes are aggregated as the representation of the synthesized sub-graph. Finally, the representations of the observed sub-graph and the synthesized sub-graph are concatenated to represent the visual relations within the complete video for classification.

These two tasks are jointly learned in an end-to-end manner to enable the synthesized sub-graph coupled with the observed sub-graph discriminative and informative for action prediction. A local graph alignment loss is proposed to constrain the anticipated visual relations of the unobserved video part as close to the corresponding realistic relations as possible. A global graph alignment loss is designed to make the grown whole graph that consists of the observed sub-graph and the synthesized sub-graph sufficiently expressive the structure information within the complete video.

The main contributions are summarized as follows:

- We propose a graph growing method to anticipate future visual relations from limited video observations for action prediction.
- We propose a joint learning of relation reasoning and relation synthesizing via combining both local and global graph alignment losses to capture the discriminative structure information.
- Extensive experiments on two video datasets demonstrate the effectiveness of the proposed method.

## Related Work

### Action Prediction

Earlier methods formulate action prediction as a probabilistic model (Ryoo 2011; Cao et al. 2013; Lan, Chen, and Savarese 2014; Li and Fu 2014). Ryoo (Ryoo 2011) calculates maximum likelihood after video segmentation to make prediction. Li et al. (Li and Fu 2014) apply probabilistic suffix tree to action prediction. Kong et al. (Kong, Kit, and Fu 2014; Kong and Fu 2016) constrain the label consistency between video segments and their corresponding full video. With the success of deep learning, many recent methods (Hu et al. 2018; Kong et al. 2018; Cai et al. 2019; Pang et al. 2019; Wang et al. 2019) resort to transferring knowledge from full videos to partial videos by using the long short-term memory (LSTM) to model the temporal information in videos. Wang et al. (Wang et al. 2019) propose a teacher-student learning based framework to transfer the action knowledge from the recognition model to the prediction model. Kong et al. (Kong et al. 2018) augment the bi-direction LSTM with a memory module to match characteristics of testing videos with training videos for action prediction. Rather than transferring the knowledge from the full video to enrich the representation of the partial video, our method forecasts the representation of the unobserved video part to generate the representation of the complete video for action prediction, thus imitating the perceptual process of human beings that innately tend to perceive the incomplete as complete.

Several other methods (Vondrick, Pirsiavash, and Torralba 2016; Pang et al. 2019; Zhao and Wildes 2019) focus on anticipating future representations from the observed video part for prediction. Vondrick et al. (Vondrick, Pirsiavash, and Torralba 2016) propose to generate visual representations of unlabeled videos to make action prediction. Pang et al. (Pang et al. 2019) predict future action features and use them to reconstruct the features of partially observed videos. Zhao et al. (Zhao and Wildes 2019) propagate residuals of features to anticipate future representations and exploit Kalman filter to make correction. Different from these methods that do not explicitly model the visual relations, our method based on the graph growing captures the spatial-temporal evolution of visual relations to synthesize the future relation representation for action prediction.

### Spatial-temporal Relation Reasoning

Spatial-temporal relation reasoning has been widely used in video understanding (Wang and Gupta 2018; Qi et al. 2018; Tsai et al. 2019; Xiao et al. 2019; Sun et al. 2019; Liu et al. 2019; Zhang et al. 2020). Qi et al. (Qi et al. 2018) exploit multi-layer perceptrons to learn the human-object interaction in videos for fine-grained action recognition. Xiao et al. (Xiao et al. 2019) introduce a dual attention mechanism and combine object attributes to represent and reason the relationship between objects and actions. Sun et al. (Sun et al. 2019) propose the relational recurrent network by combining a detection model and a recurrent neural network to jointly learn the feature extraction and relation reasoning. Several other methods build the graphs to learn different relations in videos. Zhang et al. (Zhang et al. 2020) design a multi-head temporal adjacency matrix to model various temporal relations in different granularities for action recognition. Tsai et al. (Tsai et al. 2019) regard actions as relations between visual objects and apply conditional random fields to model the video as a fully connected spatial-temporal graph to make relation reasoning. Liu et al. (Liu et al. 2019) construct three kinds of relation graphs to model the variations of human appearance, human-object interaction and human-human interaction to recognize social relationship. The aforementioned methods mainly focus on visual relation reasoning in videos. In contrast, our method does not only reason relations in the observed video part, but also synthesizes relations for the unobserved video part to predict actions.

### Our Method

Our core idea of addressing action prediction is to forecast the future spatial and temporal relations between objects from the observed video part to generate the relation representation of the complete video for action classification. The proposed graph growing method consists of a relation reasoning task that infers the visual relations from the partial observations and a relation synthesizing task that generates the visual relations of the unobserved part. In the relation reasoning, a gated graph neural network (GGNN) is utilized to perform spatial relation reasoning within video frames, and a long short-term graph convolutional network (LST-

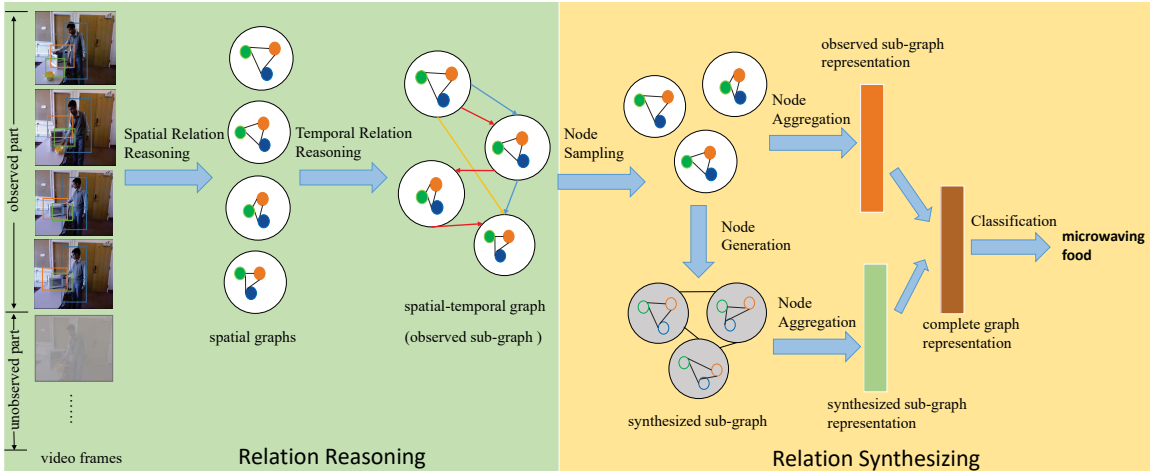


Figure 1: Overview of our method.

GCN) is proposed for multi-scale temporal relation reasoning between sequential video frames. In the relation synthesizing, a graph auto-encoder is introduced to generate spatial-temporal representation of the unobserved video part through node generation and aggregation. These two tasks are jointly learned via both local and global graph alignment losses. Figure 1 illustrates the overview of our method.

### Relation Reasoning

We build a spatial graph for each video frame, where each node represents a detected object. The node feature is extracted by making ROI Pooling on the feature map of the video frame. Each edge indicates the spatial relation between two objects and the edge feature is the representation of the union bounding boxes of the two objects. The adjacency matrix  $\mathbf{A} \in \mathbb{R}^{|V| \times |E|}$  determines how nodes communicate with each other, where  $|V|$  and  $|E|$  denote the numbers of nodes and edges in the spatial graph, respectively.

We use the gated graph neural network (GGNN) (Li et al. 2016) to perform message propagation on the spatial graph. To encourage the spatial relation reasoning in a broader region, the interaction between nodes in GGNN are replaced by the interaction between nodes and its connected edges, formulated by

$$\mathbf{a}_{v_i}^{n\top} = \mathbf{A}_{v_i}^n \left[ \mathbf{h}_{e_1}^{(n-1)} \dots \mathbf{h}_{e_{|E_s|}}^{(n-1)} \right]^\top + \mathbf{b} \quad (1)$$

where  $\mathbf{a}_{v_i}^n$  indicates the interaction between the node  $v_i$  and its directly connected edges  $\{e_1, \dots, e_{|E_s|}\}$  at the  $n$ -th timestep of propagation.  $\mathbf{A}_{v_i}^n \in \mathbb{R}^{1 \times |E_s|}$  denotes the  $i$ -th row of adjacency matrix at the  $n$ -th timestep, representing the co-occurrence relationship between the node  $v_i$  and its connected edges.  $\mathbf{h}_{e_q}^{(n-1)}$  represents the state of edge  $e_q$  at the  $(n-1)$ -th timestep of propagation.  $\mathbf{b}$  is a bias parameter. The recurrence of node propagation is performed by a gated recurrent unit (GRU). The edge feature is updated through a fully connected layer. After the spatial relation reasoning,

a graph-level representation is generated by aggregating all the node features via a soft attention mechanism:

$$\mathbf{g}_l(\mathbf{H}^N) = \tanh \left( \sum_{i=1}^{|V|} \alpha_i \tanh(\mathbf{h}_{v_i}^N) \right) \quad (2)$$

where  $\mathbf{g}_l(\mathbf{H}^N) \in \mathbb{R}^{d_v \times 1}$  ( $l = 1, \dots, L$ ) is the output graph-based representation by aggregating all the node features.  $\mathbf{H}^N = [\mathbf{h}_{v_1}^N, \dots, \mathbf{h}_{v_{|V|}}^N]$  denotes the concatenation of node features.  $\mathbf{h}_{v_i}^N$  represents the feature of node  $v_i$  after the last timestep  $N$ . We have the attention constraint:  $\sum_{i=1}^{|V|} \alpha_i = 1, \alpha_i \geq 0$ .

We build a spatial-temporal graph for the observed video part, which is called *observed sub-graph*. Each node of the spatial-temporal graph denotes a video frame and is represented by the spatial graph feature of the video frame. Each edge denotes a temporal relation between pairwise frames and is represented by a set of different scales of connections. Different connections mean temporal relations with different granularities.

To capture both long-term and short-term temporal variations in videos, we propose a long short-term graph convolutional network (LST-GCN) for multi-scale message propagation to perform temporal relation reasoning with different scales of receptive fields in the temporal domain. The multi-scale refers to sampling different numbers of video frames as nodes of the spatial-temporal graph to propagate information. It enables the spatial-temporal graph to update the information of each node and edge in both the short-term and long-term duration. Specifically, for the  $z$ -th scale,  $z$  nodes are orderly sampled with equal intervals and their features are concatenated into a vector that is used for temporal relation reasoning. A spectral convolutional neural network (Kipf and Welling 2017) is utilized for message propagation and the node features of the spatial-temporal graph are updated by

$$\mathbf{F} = \mathbf{A}_{obs} \mathbf{X}^\top \mathbf{W} \quad (3)$$

where  $\mathbf{A}_{obs} \in \mathbb{R}^{z \times z}$  represents the adjacency matrix, of which each element denotes the similarity score between two nodes. The score is computed by a softmax function that takes the concatenation of the two node features as input.  $\mathbf{X} \in \mathbb{R}^{D \times z}$  represents the updated node features after the spatial relation reasoning where  $D$  means the feature dimension.  $\mathbf{W}$  denotes a weight matrix.

After the spatial and temporal relation reasoning, for each scale, a graph-based representation  $\mathbf{E}_{obs}^z$  of the observed sub-graph is generated by feeding the concatenation of the features of the sampled nodes into the function  $\mathbf{g}_l(\cdot)$ .

## Relation Synthesizing

In the synthesizing task, a relation synthesizing graph model is proposed to make the spatial-temporal graph built in the reasoning task grow into representing the relations of the complete video. The grown graph part is named *synthesized sub-graph*. The relation synthesizing is performed in multi-scale temporal ranges to generate the unobserved video relations by preserving various structural information of the video. The nodes in the synthesized sub-graph are generated by a graph auto-encoder and then aggregated to a graph. For the  $z$ -th scale, let  $\mathbf{E}_{syn}^z$  denote the representation of the synthesized sub-graph, generated by the node aggregation defined in Eq. 2.

For the  $z$ -th scale, we firstly sample  $z$  nodes from the observed sub-graph in the same way as described in the temporal relation reasoning to generate the nodes of the synthesized sub-graph. The adjacency matrix  $\mathbf{A}_{syn}$  of the synthesized sub-graph is initialized to an identity matrix that represents every node only connects to itself. Then we use the sampled node feature  $\mathbf{F} \in \mathbb{R}^{D \times z}$  to rebuild  $\mathbf{A}_{syn}$  by two graph convolution layers:

$$\mathbf{F}' = \text{GCN}(\mathbf{F}, \mathbf{A}_{syn}) = \mathbf{D}^{-1} \mathbf{A}_{syn} \mathbf{F} \mathbf{W}_1 \quad (4)$$

$$\mathbf{Z} = \text{GCN}(\mathbf{F}', \mathbf{A}_{syn}) = \mathbf{D}^{-1} \mathbf{A}_{syn} \mathbf{F}' \mathbf{W}_2 \quad (5)$$

where  $\mathbf{D}$  is the degree matrix of  $\mathbf{A}_{syn}$ .  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are trainable weight matrices.  $\mathbf{Z}$  is a mean vector matrix. Then we employ inner product decoder to update the adjacency matrix by  $\hat{\mathbf{A}}_{syn} = \varphi(\mathbf{Z}\mathbf{Z}^T)$ , where  $\varphi(\cdot)$  is a Sigmoid function. After updating the adjacency matrix, the generated  $z$  nodes in the synthesized sub-graph are formulated by

$$\mathbf{F}^* = \text{GCN}(\mathbf{F}, \hat{\mathbf{A}}_{syn}). \quad (6)$$

The rebuilt  $\hat{\mathbf{A}}_{syn}$  preserves the temporal relations in the observed sub-graph and the features of the sampled nodes capture spatial relations in video frames. Thus the generated  $z$  nodes maintain the structural information of the observed sub-graph over both space and time. Finally, the graph-level representation  $\mathbf{E}_{syn}^z$  of the synthesized sub-graph at the  $z$ -th scale is calculated by the node aggregation defined in Eq. 2.

## Joint Learning

To enable the representation of the synthesized sub-graph sufficiently realistic for action classification, we propose a

local graph alignment loss, given by

$$Loss_{local} = \sum_{i=1}^I \sum_{z=1}^Z \|\mathbf{E}_{syn,i}^z \odot \omega - \mathbf{E}_{gts,i}^z \odot \omega\|_F^2 \quad (7)$$

where  $\mathbf{E}_{syn,i}^z \in \mathbb{R}^{d \times z}$  and  $\mathbf{E}_{gts,i}^z \in \mathbb{R}^{d \times z}$  respectively indicate the representations of the synthesized sub-graph and the ground-truth sub-graph of the unobserved video part for the  $i$ -th training video at the  $z$ -th scale.  $\omega$  is a weight vector and  $\odot$  represents element-level multiplication.

With the observed sub-graph representation  $\mathbf{E}_{obs,i}^z$  and the synthesized sub-graph representation  $\mathbf{E}_{syn,i}^z$  for the  $i$ -th training video at the  $z$ -th scale, we have a complete graph representation  $\mathbf{E}_{gen,i}^z = \mathbf{g}_l([\mathbf{E}_{obs,i}^z, \mathbf{E}_{syn,i}^z])$  of the  $i$ -th training video at the  $z$ -th scale.

To make the synthesized graph representation contain global structure information for generating the spatial-temporal relations of the fully observed video, we propose a global alignment loss, formulated by

$$Loss_{global} = \sum_{i=1}^I \left\| \frac{1}{Z} \sum_{z=1}^Z \phi(\mathbf{E}_{gen,i}^z) - \frac{1}{Z} \sum_{z=1}^Z \phi(\mathbf{E}_{gt,i}^z) \right\|_F^2 \quad (8)$$

where  $\mathbf{E}_{gt,i}^z$  indicates the representation of the ground-truth graph of the complete video for the  $i$ -th training video at the  $z$ -th scale.  $\phi$  denotes a mapping function that maps a graph representation feature to the Reproducing Kernel Hilbert Space (RKHS). In this paper, we use the Gaussian kernel:

$$k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma}}.$$

## Loss Function

The generated complete graph representation at each scale is used to produce the action category probability. All the action category probabilities from all the scales are summed up and fed into a softmax function to output the final action category label. We use a cross-entropy loss to train the classifier, given by

$$Loss_{cls} = \sum_{i=1}^I [-y_i \log \hat{y}_i - (1 - y_i)(1 - \hat{y}_i)] \quad (9)$$

where  $y_i$  and  $\hat{y}_i$  are the ground-truth label and predicted label of the  $i$ -th training video, respectively.

Therefore, the relation reasoning task and the relation synthesizing task are jointly learned through the whole loss function:

$$Loss = Loss_{cls} + \lambda_1 Loss_{local} + \lambda_2 Loss_{global} \quad (10)$$

where  $\lambda_1$  and  $\lambda_2$  represent the balance parameters. In the training stage, we firstly learn the relation reasoning task using Eq. 9 and then jointly learn the relation reasoning and synthesizing tasks using Eq. 10.

## Experiments

### Datasets

We conduct experiments to evaluate our method on two datasets: UCF101 (Soomro, Zamir, and Shah 2012) and 20BN-something-something (Goyal et al. 2017).

Feature	Method	Observation ratio					
		0.1	0.2	0.3	0.5	0.7	0.9
ResNet-18	MSSC(Cao et al. 2013)	34.05	–	58.32	62.52	63.55	62.67
	MTSSVM(Kong, Kit, and Fu 2014)	40.05	–	80.02	82.13	82.49	83.18
	DeepSCN(Kong, Tao, and Fu 2017)	45.02	–	82.19	84.92	85.59	86.02
	mem-LSTM(Kong et al. 2018)	51.02	–	86.75	88.37	89.22	89.97
	MSRNN(Hu et al. 2016)	68.01	–	<b>88.71</b>	<b>89.25</b>	<b>89.92</b>	<b>90.23</b>
	<b>ours</b>	<b>75.85</b>	<b>81.72</b>	87.78	88.69	89.42	90.15
ResNet-50	Context-aware + loss in (Jain et al. 2016)	–	30.60	–	71.10	–	–
	Context-aware + loss in (Ma, Sigal, and Sclaroff 2016)	–	22.60	–	73.10	–	–
	MS-LSTM (Hu et al. 2016)	–	80.50	–	83.40	–	–
	AA-GAN (Gammulle et al. 2019)	–	84.20	–	85.60	–	–
	<b>ours</b>	<b>84.11</b>	<b>88.50</b>	<b>89.80</b>	<b>90.90</b>	<b>91.40</b>	<b>91.80</b>
InceptionV4	RGN-KF(Zhao and Wildes 2019)	<b>83.30</b>	85.16	87.78	<b>91.50</b>	92.03	92.85
	AAPNet(Kong, Tao, and Fu 2020)	59.85	80.44	87.12	86.65	88.34	90.92
	<b>ours</b>	82.36	<b>85.57</b>	<b>88.97</b>	91.32	<b>92.41</b>	<b>93.02</b>

Table 1: Accuracies (%) of different methods on the UCF101 dataset using ResNet-18, ResNet-50 and InceptionV4 features

method	observation ratio					
	0.1	0.2	0.3	0.5	0.7	0.9
mem-LSTM(Kong et al. 2018)	14.92	17.16	18.08	20.44	23.22	24.46
MS-LSTM(Sadegh Aliakbarian et al. 2017)	16.89	16.57	16.82	16.71	16.95	17.08
MSRNN(Hu et al. 2016)	20.62	20.47	21.02	22.45	24.05	27.13
<b>ours</b>	<b>21.17</b>	<b>21.49</b>	<b>23.30</b>	<b>27.68</b>	<b>30.23</b>	<b>30.55</b>

Table 2: Accuracies (%) of different methods on the 20BN-something-something dataset.

The UCF101 dataset has been widely used for action recognition and prediction because of its wide range of activities. It consists of 13,320 videos covering 101 action categories. The actions in this dataset contain five types, including human-object interaction, human-human interaction, body-motion only, playing musical instruments and sports. This dataset provides three official splits for training and validation and we report the average accuracy over the three splits by following the standard practice.

The 20BN-something-something dataset is a large collection of 108,499 densely-labeled video clips across 174 labels. These collected videos show more fine-grained actions of human with everyday objects in real life, thus recognizing them requires a detailed understanding of actions and scenes. We use the standard and official subset that contains 21 action categories, including “Opening something”, “Closing something”, “Turning something upside down”, “Pretending to turn something upside down” and so on. There are 11,101 short videos for training and 1,568 videos for validation. We report the results by averaging classification accuracies over all the classes.

## Implementation Details

**Feature Representation.** For the UCF101 dataset, we employ a Faster R-CNN model pre-trained on the ImageNet-1k dataset to detect objects in video frames. Then we extract three kinds of Two-Stream CNN features to represent the objects and their relations: ResNet18 (He et al. 2016), ResNet50 (He et al. 2016) and InceptionV4 (Wang et al. 2016). For the ResNet18 feature, we train two net-

works for the RGB and optical flow streams, respectively, following (Hu et al. 2018). For the RGB stream, we finetune ResNet-18 pretrained on ImageNet. For the optical flow stream, we train ResNet-18 from scratch. For the ResNet50 feature, we finetune ResNet-50 pretrained on ImageNet for both RGB and optical flow streams, following (Gammulle et al. 2019). For the InceptionV4 feature (Wang et al. 2016), we use TRN (Wang et al. 2016) trained on Kinetics (Kay et al. 2017) with BN-Inception (Ioffe and Szegedy 2015) as the backbone and finetune it for both GB and optical flow streams, following (Zhao and Wildes 2019). For all the three kinds of features, we use the feature map from the last convolutional layer and make ROI Pooling to extract the node and edge features.

For the 20BN-something-something dataset, we train a Faster R-CNN model with the ResNet-50-FPN backbone (Lin et al. 2017). Since the ground-truth bounding boxes of objects are not available on this dataset, we manually annotate bounding boxes of objects in 10 frames sampled from each video and nearly 2,000 videos (about 20% of the training videos) are annotated. The objects are extracted by the trained Faster R-CNN model and the threshold of Intersection-over-Union(IoU) for proposals in non-maximum suppression is set to 0.5. We extract the features of bounding boxes from the last fully connected layer of the Faster R-CNN model. In the spatial graph, the nodes are initialized by the features of the detected object bounding boxes and the edges are initialized by the union bounding boxes of the two corresponding objects.

**Model Setting.** In the relation reasoning, the unit number

method	UCF101						20BN-something-something					
	observation ratio						observation ratio					
	0.1	0.2	0.3	0.5	0.7	0.9	0.1	0.2	0.3	0.5	0.7	0.9
w/o rea & syn	70.95	72.20	72.20	72.78	73.07	72.90	17.16	17.79	18.73	20.33	21.94	22.68
w/o synthesizing	78.58	82.27	82.26	82.64	86.37	86.80	14.80	19.20	20.22	24.36	27.74	28.06
w/o reasoning	76.63	81.22	83.60	85.52	86.30	86.80	19.98	19.32	20.11	22.06	25.32	27.28
w/o local loss	70.90	81.37	83.63	85.50	86.34	86.81	19.71	20.17	20.28	23.17	25.68	26.79
w/o global loss	74.90	85.30	88.50	90.70	91.80	92.02	18.43	18.75	20.16	22.83	24.75	24.40
<b>ours</b>	<b>82.36</b>	<b>85.57</b>	<b>88.97</b>	<b>91.32</b>	<b>92.41</b>	<b>93.02</b>	<b>21.17</b>	<b>21.49</b>	<b>23.30</b>	<b>27.68</b>	<b>30.23</b>	<b>30.55</b>

Table 3: Accuracies (%) of ablation studies on the UCF101 and 20BN-something-something datasets.

method	UCF101						20BN-something-something					
	observation ratio						observation ratio					
	0.1	0.2	0.3	0.5	0.7	0.9	0.1	0.2	0.3	0.5	0.7	0.9
I-Matrix	80.05	84.93	86.16	87.22	87.92	88.40	17.16	18.11	18.73	20.33	21.94	22.68
S-Matrix ( <b>ours</b> )	<b>82.36</b>	<b>85.57</b>	<b>88.97</b>	<b>91.32</b>	<b>92.41</b>	<b>93.02</b>	<b>21.17</b>	<b>21.49</b>	<b>23.30</b>	<b>27.68</b>	<b>30.23</b>	<b>30.55</b>

Table 4: Accuracies (%) of different relation matrix  $\mathbf{A}$  on the UCF101 and 20BN-something-something datasets.

of GRU layers is set to 512 and the number of propagation is set to 3. The feature dimensions of both initial nodes and edges are reduced to 512 by a linear layer. In the relation synthesizing, the dimension of graph representation feature is set to 512. We use two graph convolutional layers to rebuild the adjacency matrix  $\mathbf{A}_{syn}$  and one graph convolutional layer to refresh the node feature  $\mathbf{F}$ . The number of scale is set to 5 for 20BN-something-something and 8 for UCF101. The balance parameter  $\lambda_1$  is set to 0.125 and  $\lambda_2$  is set to 1. We randomly split 10% of the training set as a validation set. All the networks are trained from scratch with an initial learning rate of 0.00005. The Adam optimizer (Kingma and Ba 2015) and the SGD optimizer are employed with a batch size of 48 for optimization. We use PyTorch 0.4.1 and train the model for 500 epochs on one GTX-1080Ti GPU.

### Comparison with State-of-the-art Methods

To evaluate the effectiveness of our method, we compare our method with several state-of-the-art methods and report the action prediction accuracies at the observation ratios of  $\{0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$ .

Table 1 shows the comparison results on the UCF101 dataset. The results of all the compared methods are directly copied from their original papers. For fair comparison, we use the same visual features (i.e., ResNet-18, ResNet-50 and Inception V4) as the other methods. From Table 1, it is noteworthy to make several observations. First, when using the ResNet-18 feature, our method achieves better performance with a gain of 7% at the observation ratio of 0.1 and comparable performance at the other observation ratios, which validates the benefit of modeling the spatial-temporal evolution of visual relations to the early prediction. Second, when using the ResNet-50 feature, our method achieves the best results at all the observation ratios. The closest to our method is AA-GAN that generates visual representation of the unobserved video part via GAN. Our method outperforms it by 4% and 5% at the observation ratios of 0.2 and 0.5, respec-

tively, clearly demonstrating the superiority of anticipating future visual relations via graph growing. Third, when using the Inception V4 feature, our method outperforms the state-of-the-art methods for most observation ratios.

Table 2 shows the comparison results on the 20BN-something-something dataset. All the methods use the same feature for fair comparison. For the compared methods, the extracted features of bounding boxes are concatenated to represent each video frame. It can be observed that the performance of our method is superior to that of other methods, which demonstrates that reasoning observed relations and anticipating future relations is beneficial to improving the action prediction accuracy.

### Ablation Studies

**Evaluation on Each Component.** Table 3 reports the ablation study results of different individual components on both UCF101 and 20BN-something-something datasets. “w/o rea & syn” represents the model without relation reasoning and relation synthesizing, which concatenates the features of the detected object bounding boxes and the features of the union bounding boxes of two objects as input, and directly uses the cross-entropy loss for training. “w/o synthesizing” means only performing relation reasoning from the observed video part without forecasting future relations. “w/o reasoning” means performing synthesizing future with the observed video part directly without performing relation reasoning. We can observe that when removing the relation reasoning or the relation synthesizing, the results will substantially degrade at all the observation ratios, which validates that both of them are critical to the prediction performance.

**Evaluation on Different Losses.** To evaluate how the local and global graph alignment losses affect the prediction performance, we conduct experiments with different losses on both UCF101 and 20BN-something-something datasets. “w/o local loss” represents only using the global graph





















Observation Ratio	10%	30%	50%	70%	100%
<b>Picking sth up</b>					
<b>Only reasoning</b>	Turning sth upside down	Turning sth upside down	Putting sth upright on the table	Picking sth up	Picking sth up
<b>Ours</b>	Turning sth upside down	Turning sth upside down	Picking sth up	Picking sth up	Picking sth up
<b>Pretending to sprinkle air onto sth</b>					
<b>Only reasoning</b>	Pretending to open sth without actually opening it	Pretending to open sth without actually opening it	Pretending to open sth without actually opening it	Pretending to sprinkle air onto sth	Pretending to sprinkle air onto sth
<b>Ours</b>	Pretending to open sth without actually opening it	Pretending to open sth without actually opening it	Pretending to sprinkle air onto sth	Pretending to sprinkle air onto sth	Pretending to sprinkle air onto sth
<b>Putting sth into sth</b>					
<b>Only reasoning</b>	Turning sth upside down	Turning sth upside down	Turning sth upside down	Turning sth upside down	Putting sth into sth
<b>Ours</b>	Turning sth upside down	Turning sth upside down	Putting sth into sth	Putting sth into sth	Putting sth into sth
<b>Closing sth</b>					
<b>Only reasoning</b>	Opening sth	Opening sth	Stuffing sth into sth	Stuffing sth into sth	Stuffing sth into sth
<b>Ours</b>	Opening sth	Closing sth	Closing sth	Closing sth	Closing sth

Figure 2: Prediction examples on the 20BN-something-something dataset. The ground-truth labels in green are given on the left side and predicted labels at different observation ratios are under the video frames. The labels in blue represent correctly predicted labels and the labels in red represent wrongly predicted labels.

alignment loss and the classification loss for training. “w/o global loss” represents only using the local graph alignment loss and the classification loss for training. From the result in Table 3, we can observe that the two losses work together to make a positive impact on the prediction performance.

**Evaluation on Different Relation Matrices A.** We compare performances of different relation matrices  $\mathbf{A}$  in the spatial graph, as shown in Table 4. “I-Matrix” means the elements in relation matrix  $\mathbf{A}$  that represent nodes associated with edges are initialized to 1 and “S-Matrix” means the elements in relation matrix  $\mathbf{A}$  that represent the nodes associated with edges is initialized by formulating the dot similarity between the node pairs. In this paper, we initialize  $\mathbf{A}$  using dot similarity between node pairs to achieve better results.

### Qualitative Analysis

To qualitatively analyze how the relation synthesizing affect the prediction results, we show several exemplars of predicted action category labels at different observations ratios on the 20BN-something-something dataset in Figure 2, where “Only reasoning” represents only performing relation reasoning from the observed video frames without anticipating future relations. It is interesting to observe that our

method is able to make accurate predictions earlier than the model that only performs relation reasoning, which clearly demonstrates the importance of synthesizing the relations of the unobserved video frames to generate the complete relation representation for action prediction.

## Conclusion

We have presented a spatial-temporal graph growing method for action prediction from partial videos. A relation reasoning model based on a gated graph neural network and a long short-term graph convolutional network have been designed to infer the spatial and temporal relations between visual objects of the observed video part. A relation synthesizing model based on a graph auto-encoder has been proposed to generate node and edge representations for the unobserved video part. The relation reasoning and synthesizing models are jointly learned via an integration of the local and global graph alignment losses. Our method can successfully interpret the observed video content and anticipate the future video representation with fine-grained relations to make prediction decisions. Experiments have shown the superior performance of our method.

## Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants No. 62072041 and No. 61673062.

## References

- Cai, Y.; Li, H.; Hu, J.-F.; and Zheng, W.-S. 2019. Action Knowledge Transfer for Action Prediction with Partial Videos. In *AAAI Conference on Artificial Intelligence*, 8118–8125.
- Cao, Y.; Barrett, D.; Barbu, A.; Narayanaswamy, S.; Yu, H.; Michaux, A.; Lin, Y.; Dickinson, S.; Mark Siskind, J.; and Wang, S. 2013. Recognize Human Activities from Partially Observed Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2658–2665.
- Ehrenstein, W. H.; Spillmann, L.; and Sarris, V. 2003. Gestalt Issues in Modern Neuroscience. *Axiomathes* 13(3-4): 433–458.
- Farahzad, F. 1998. A Gestalt Approach to Manipulation. *Perspectives: Studies in Translatology* 6(2): 153–158.
- Gammulle, H.; Denman, S.; Sridharan, S.; and Fookes, C. 2019. Predicting the Future: A Jointly Learnt Model for Action Anticipation. In *IEEE International Conference on Computer Vision*, 5561–5570.
- Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Freund, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The “Something Something” Video Database for Learning and Evaluating Visual Common Sense. In *IEEE International Conference on Computer Vision*, 5843–5851.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Hu, J.-F.; Zheng, W.-S.; Ma, L.; Wang, G.; and Lai, J. 2016. Real-time RGB-D Activity Prediction by Soft Regression. In *European Conference on Computer Vision*, 280–296.
- Hu, J.-F.; Zheng, W.-S.; Ma, L.; Wang, G.; Lai, J.; and Zhang, J. 2018. Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(11): 2568–2583.
- Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, 448–456.
- Jain, A.; Singh, A.; Koppula, H. S.; Soh, S.; and Saxena, A. 2016. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *Robotics and Automation, 2016 IEEE International Conference*, 3118–3125. IEEE.
- Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.
- Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations*.
- Kong, Y.; and Fu, Y. 2016. Max-margin Action Prediction Machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(10): 1844–1858.
- Kong, Y.; Gao, S.; Sun, B.; and Fu, Y. 2018. Action Prediction from Videos via Memorizing Hard-to-Predict Samples. In *AAAI Conference on Artificial Intelligence*, 7000–7007.
- Kong, Y.; Kit, D.; and Fu, Y. 2014. A Discriminative Model with Multiple Temporal Scales for Action Prediction. In *European Conference on Computer Vision*, 596–611.
- Kong, Y.; Tao, Z.; and Fu, Y. 2017. Deep Sequential Context Networks for Action Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3662–3670.
- Kong, Y.; Tao, Z.; and Fu, Y. 2020. Adversarial Action Prediction Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(3): 539–553.
- Lan, T.; Chen, T.-C.; and Savarese, S. 2014. A Hierarchical Representation for Future Action Prediction. In *European Conference on Computer Vision*, 689–704.
- Li, K.; and Fu, Y. 2014. Prediction of Human Activity by Discovering Temporal Sequence Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(8): 1644–1657.
- Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. S. 2016. Gated Graph Sequence Neural Networks. In *4th International Conference on Learning Representations*.
- Lin, T.-Y.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 936–944.
- Liu, X.; Liu, W.; Zhang, M.; Chen, J.; Gao, L.; Yan, C.; and Mei, T. 2019. Social Relation Recognition from Videos via Multi-scale Spatial-temporal Reasoning. In *IEEE conference on Computer Vision and Pattern Recognition*, 3566–3574.
- Ma, S.; Sigal, L.; and Sclaroff, S. 2016. Learning Activity Progression in Lstms for Activity Detection and Early Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1942–1950.
- Pang, G.; Wang, X.; Hu, J.-F.; Zhang, Q.; and Zheng, W.-S. 2019. DBDNet: Learning Bi-directional Dynamics for Early Action Prediction. In *International Joint Conference on Artificial Intelligence*, 897–903.
- Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *European Conference on Computer Vision*, 407–423.
- Ryoo, M. S. 2011. Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos. In



*IEEE International Conference on Computer Vision*, 1036–1043.

Sadegh Aliakbarian, M.; Sadat Saleh, F.; Salzmman, M.; Fernando, B.; Petersson, L.; and Andersson, L. 2017. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*, 280–289.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv preprint arXiv:1212.0402*.

Spillmann, L. 2006. From Perceptive Fields to Gestalt. *Progress in Brain Research* 155(1): 67–92.

Sun, C.; Shrivastava, A.; Vondrick, C.; Sukthankar, R.; Murphy, K.; and Schmid, C. 2019. Relational Action Forecasting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 273–283.

Tsai, Y.-H. H.; Divvala, S.; Morency, L.-P.; Salakhutdinov, R.; and Farhadi, A. 2019. Video Relationship Reasoning Using Gated Spatio-temporal Energy Graph. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10424–10433.

Vondrick, C.; Pirsivash, H.; and Torralba, A. 2016. Anticipating Visual Representations from Unlabeled Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 98–106.

Wang, L.; Li, W.; and Van Gool, L. 2018. Appearance-and-Relation Networks for Video Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1430–1439.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision*, 20–36.

Wang, X.; and Gupta, A. 2018. Videos as Space-time Region Graphs. In *European Conference on Computer Vision*, 413–431.

Wang, X.; Hu, J.-F.; Lai, J.-H.; Zhang, J.; and Zheng, W.-S. 2019. Progressive Teacher-Student Learning for Early Action Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3556–3565.

Xiao, T.; Fan, Q.; Gutfreund, D.; Monfort, M.; Oliva, A.; and Zhou, B. 2019. Reasoning About Human-Object Interactions Through Dual Attention Networks. In *IEEE International Conference on Computer Vision*, 3918–3927.

Zhang, J.; Shen, F.; Xu, X.; and Shen, H. T. 2020. Temporal Reasoning Graph for Activity Recognition. *IEEE Transactions on Image Processing* 29(4): 5491–5506.

Zhao, H.; and Wildes, R. P. 2019. Spatiotemporal Feature Residual Propagation for Action Prediction. In *IEEE International Conference on Computer Vision*, 7002–7011.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal Relational Reasoning in Videos. In *European Conference on Computer Vision*, 831–846.