Multimodal Global Relation Knowledge Distillation for Egocentric Action Anticipation

Yi Huang[1,2], Xiaoshan Yang[1,2,3], Changsheng Xu[1,2,3,4*]
[1]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Peng Cheng Lab, Shenzhen, 518055, China
[4]CASIA-LLVision Joint Lab, China
{yi.huang,xiaoshan.yang,csxu}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we consider the task of action anticipation on egocentric videos. Previous methods ignore explicit modeling of the global context relation among past and future actions, which is not an easy task due to the vacancy of unobserved videos. To solve this problem, we propose a Multimodal Global Relation Knowledge Distillation (MGRKD) framework to distill the relation knowledge learned from full videos to improve the action anticipation task on partially observed videos. The proposed MGRKD has a teacher-student learning strategy, where either the teacher or student model has three branches of global relation graph networks (GRGN) to explore the pairwise relations between past and future actions based on three kinds of features (*i.e.*, RGB, motion or object). The teacher model has a similar architecture with the student model, except that the teacher model uses true feature of the future video snippet to build the graph in GRGN while the student model uses a progressive GRU to predict an initialized node representation of future snippet for reasoning on GRGN. Through the teacher-student learning strategy, the discriminative features and privileged relation knowledge of the past and future actions learned in the teacher model can be distilled to the student model. We perform experiments on two egocentric video datasets EPIC-Kitchens and EGTEA Gaze+. The results show that the proposed framework achieves state-of-the-art performances.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Temporal reasoning**.

KEYWORDS

egocentric action anticipation, graph network, knowledge distillation

*Corresponding author

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475327

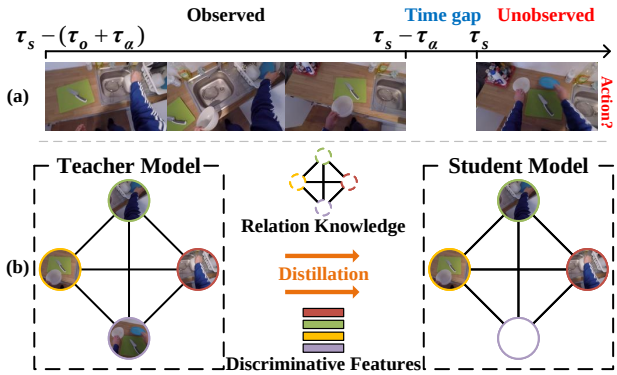

Figure 1: (a) Illustration of action anticipation task: predicting the action starting at τ_s by observing a video starting at $\tau_s - (\tau_o + \tau_\alpha)$ and ending at $\tau_s - \tau_\alpha$, where τ_o is the length of observed video and τ_α is the anticipation time gap. (b) Our main idea: distilling the discriminative features and global relation knowledge of teacher model learned from full videos into the student model for action anticipation.

ACM Reference Format:
Yi Huang, Xiaoshan Yang, and Changsheng Xu. 2021. Multimodal Global Relation Knowledge Distillation for Egocentric Action Anticipation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475327

1 INTRODUCTION

Action anticipation, *i.e.*, predicting an action *before it actually begins*, is a fundamental ability for human beings to make decisions when interacting with the environment. Formally, as shown in Figure 1(a), this task is defined as predicting the action of a video starting at time τ_s by observing the video starting at $\tau_s-(\tau_o+\tau_\alpha)$ and ending at $\tau_s-\tau_\alpha$, where τ_o is the length of observed video and τ_α is the gap of anticipation time. The video content after time $\tau_s-\tau_\alpha$ is unavailable when anticipating. It is worth noting that the observed videos may have different actions from the unobserved ones. Although this task is firstly proposed in third-person videos [22, 24, 25, 42], anticipating an action from egocentric (first-person) videos has also attracted much attention since it has many real-world applications, such as autonomous vehicles [1, 31], human-robot interaction [22, 37] and wearable assistants [19, 40]. For example, the ability of an intelligent wearable system for predicting what action the wearer will perform

in the near future is critical to prepare assistance in advance for the wearer to execute this action. The central challenge of egocentric action anticipation arises in inferring the future by creating connection between past and future events based on partially observed videos. However, due to the uncertainty of the future, it is difficult to directly utilize current advanced action recognition models, such as Temporal Segment Network [43] and Two-Stream CNNs [39], for action anticipation. Because most of these methods mainly focus on capturing the discriminative visual and motion content of the observed videos which is not sufficient to reason out future actions.

Recently, there are some successful attempts to address the challenge of egocentric action anticipation. For example, Miech *et al.* [32] propose a basic model to predict current action based on observed videos, and then build a transitional model to predict future action based on the current action labels. Furnari *et al.* [9] utilize two LSTMs to handle disentangled sub-tasks of summarizing past observations and anticipating future actions based the hidden vectors learned in the observed videos. Zhang *et al.* [52] develop [9] by applying textual pre-training to acquire tacit knowledge to solve the visual gap problem. Besides, the *action* is divided into the form of *verb* and *noun* to implement analysis-based prediction. Although these methods have achieved significant progress in egocentric action anticipation, they always predict the future action based on an integrated state vector which represents all the complex information in the observed videos, such as objects, motions and the relations among them. This kind of architecture will limit the anticipation performances because the integrated representation may ignore a potential strong connection between past and future actions with long time intervals. For example, if a man has just performed actions of "*stand up*" and "*open fridge*", the future action can be "*take eggs*", "*take fruits*", "*take milk*", etc. Based on the integrated representation of the primary objects and motions contained in the observed videos, it is not easy to decide what is the correct future action. In contrast, if we can identify that the person holds a cup of coffee, the correct result "*take milk*" can be easily obtained due to the strong prior dependency between "*coffee*" and "*milk*". This kind of relation can be captured by globally considering the pairwise relations between past and future actions.

To this end, we know that comprehensively exploring the global context relations between past and future actions is important in action anticipation. Inspired by the success of graph networks in modeling the structure information [20, 48, 50], it is straightforward to consider using graph networks to capture the global relations. However, the vacancy of the unobserved video brings extra difficulties because we can not directly construct the reasoning graph based on unknown nodes. **To solve this problem, we propose an end-to-end graph-based knowledge distillation framework, named as Multimodal Global Relation Knowledge Distillation (MGRKD).** As shown in Figure 1(b), our work is designed to distill the discriminative features and the global context relation knowledge learned in the full videos to improve the action anticipation task on partially observed videos.

The proposed MGRKD adopts a teacher-student learning strategy. **The student model** is designed to simulate the real inference environment, where only the observed videos are available for action anticipation. We build three branches of global relation graph networks (GRGN) to explore the pairwise relations between the past and future actions. Each GRGN uses one of three kinds of features (*i.e.*, RGB, motion or object) to create the input graph, where the node represents the feature of a video snippet and the edge denotes the relation between two video snippets. For the convenience of graph reasoning, a progressive GRU network is adopted to predict an initialized representation of future video snippet. With the GRGN, we can reason out the discriminative feature of the future snippet, which will be further used to predict future action class. Moreover, the predicted results obtained by three branches of GRGNs are combined by late fusion strategy to exploit the complementarity of different kinds of features. **The teacher model** has a similar architecture with the student model, except that the true feature of the unobserved video snippet is used to build the graph. The teacher model practically solves an easier task of action recognition based on full videos. Through the teacher-student learning strategy, the discriminative features and privileged relation knowledge of the past and future actions learned in the teacher model can be distilled to the student model during train phase. At test phase, only the student model is retained to anticipate the future action. We conduct experiments on two large-scale egocentric video datasets, EPIC-Kitchens [5] and EGTEA Gaze+ [26]. The results demonstrate that the proposed method achieves the state-of-the-art performances on the action anticipation task.

In summary, the main contributions of this work are three-fold:

- We propose to model the global pairwise relations between past and future actions with graph networks, which can effectively capture useful dependencies with long time intervals for action anticipation.
- We design a novel multimodal global relation knowledge distillation framework to distill the discriminative features and global context relations learned from the full videos to improve the action anticipation on partially observed videos. To the best of our knowledge, this is the first work of relation distillation in the action anticipation task.
- We evaluate the proposed MGRKD on two large-scale egocentric video datasets (*i.e.*, EPIC-Kitchens [5] and EGTEA Gaze+ [26]). Extensive experimental results demonstrate that the proposed method achieves state-of-the-art performances.

2 RELATED WORK

Egocentric Action Anticipation. The task of action anticipation aims at predicting an action before it actually begins [9, 11]. This definition distinguishes the anticipation task from early action recognition [15, 21, 45]. Compared with action anticipation in third-person vision [22, 24, 25, 42], fewer works focus on the egocentric action anticipation task [7, 32]. Miech *et al.* [32] propose a model based on Markov processing to establish transition relation between past and future actions. Furnari *et al.* [7] consider the action anticipation as a multi-label classification problem since the the future action is uncertain, and study the design of loss functions. As an improvement, Camporese *et al.* [2] extend the idea of label smoothing by extracting semantics from the target labels and distill the semantic information into the model during training. Furnari *et al.* [9] utilize two LSTMs to handle disentangled sub-tasks of summarizing past observations and anticipating future actions based the hidden vectors learned in the observed videos. In order to bridge

visual gap between past and future, Zhang *et al.* [52] adopt multi-modal information, including both visual features of videos and sequential text instructions of actions, to perform action anticipation via integrating intuition and analysis. Liu *et al.* [27] find that hand movement reveals critical information about the future activity in egocentric vision, and propose to jointly predict the egocentric hand motion, interaction hotspots and future action. Qi *et al.* [36] propose a self-regulated learning framework to regulate model consecutively to produce representation that emphasizes the novel information in the frame of the current time-stamp in contrast to previously observed content. Wu *et al.* [47] decompose the action anticipation into a series of future feature predictions by using contrastive learning to pick the correct future states from the given observed video features and future true features. Most existing methods use an integrated state vector to capture all the complex information in the observed videos, which may limit the anticipation performances. In contrast, our method explores the global context relations between past and future actions.

Graph Convolutional Networks. After proposed in [20], graph convolution networks have shown convincing successes in modeling relations and have been widely applied in multiple computer vision and multimedia tasks, such as zero/few-shot image classification [3, 18, 46], visual question answering [33] and visual captioning [49]. Similarly, many works have been proposed to analyze videos [10, 12, 16, 34, 35, 44, 51, 53]. For example, Zeng *et al.* [51] propose to use GCNs to model the relation of video clips to further improve the localization accuracy. Pan *et al.* [34] propose a spatio-temporal graph to model the interactions of objects detected in video to improve the captioning performance. To the best of our knowledge, we are the first to apply GCNs in action anticipation task to explore the relations among past and future actions.

Knowledge Distillation. Knowledge distillation is first proposed in [14] to distill the knowledge from a large model into a small model by minimizing the KL divergence between their logits distributions. Due to its simplicity and effectiveness, knowledge distillation has been widely used in model compression [23, 29, 41]. Laterly, Lopez-Paz *et al.* [28] generalize distillation to incorporate privileged information, which is available during training but not accessible during testing. Gupta *et al.* [13] treat the extra modality as the privileged information for cross-modal distillation. Zhou *et al.* [54] propose to use image-text matching model to distill word-region alignment information for image captioning. Pan *et al.* [34] propose to distill object interaction information from a spatio-temporal graph for video captioning. Camporese *et al.* [2] generalize the label smoothing idea by extrapolating semantic priors from the action labels. The work more related to ours is [45], where full video is regarded as privileged information to be distilled into the early action prediction model. It is worth noting that this method only distills the features from full video. Differently, our method can distill the global context relations between past and future actions learned in full video to improve the performance of action anticipation task.

3 PROPOSED APPROACH

3.1 Framework Overview

The egocentric action anticipation task, as defined in [8, 9], aims to predict the action label y of a video starting at time τ_s by observing

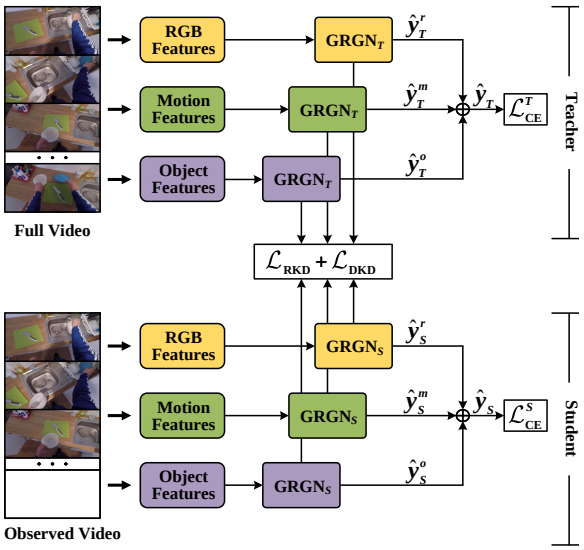

Figure 2: An overview of the proposed Multimodal Global Relation Knowledge Distillation (MGRKD). r_i, m_i and o_i are input RGB, motion and object features. GRGN is the proposed global relation graph network. In either the student model or the teacher model, results predicted by the classifiers based the outputs of three GRGNs are combined by late fusion strategy to anticipate the future action.

the video starting at $\tau_s - (\tau_o + \tau_\alpha)$ and ending at $\tau_s - \tau_\alpha$, where τ_o is the length of observed video and τ_α is the time gap for anticipation. For simplicity, as in [9], we segment the video into snippets, where each of them has δ seconds. With this scheme, we get l video snippets $\{V_1, V_2, \ldots, V_l\}$ during time $[\tau_s - (\tau_o + \tau_\alpha), \tau_s - \tau_\alpha]$ and $a-1$ video snippets $\{V_{l+1}, V_{l+2}, \ldots, V_{l+a-1}\}$ during the anticipation gap $[\tau_s - \tau_\alpha, \tau_s]$. Moreover, we use V_{l+a} to denote the video snippet after τ_s that need to be anticipated. The challenge of this task lies largely in that $\{V_{l+1}, V_{l+2}, \ldots, V_{l+a}\}$ are unavailable when anticipating.

To comprehensively capture the important visual content (*e.g.*, objects and motions) contained in the videos, we extract three kinds of features including RGB feature, motion feature and object feature to construct our framework for action anticipation. As in [8, 9], for each video snippet V_i, we extract RGB features $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_l\}$ via pretrained spatial CNNs, motion features $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_l\}$ via motion CNNs which can process optical flow information, and object features $\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_l\}$ via a pretrained object detector. More details of the feature extraction networks are illustrated in Section 4.2.

After obtaining the three kinds of features, as shown in Figure 2, we build a multi-modal knowledge distillation framework which implements a teacher-student learning strategy. Either the teacher model or the student model has three branches. Each branch is a Global Relation Graph Network (GRGN) which explores the global context relations between the past and future actions based on one kind of features (*i.e.*, RGB, motion or object). For the student model, only observed video snippets are available to build the relation graph. Therefore, we initialize the nodes of future video snippets by a progressive GRU. With the GRGN, we can reason out the

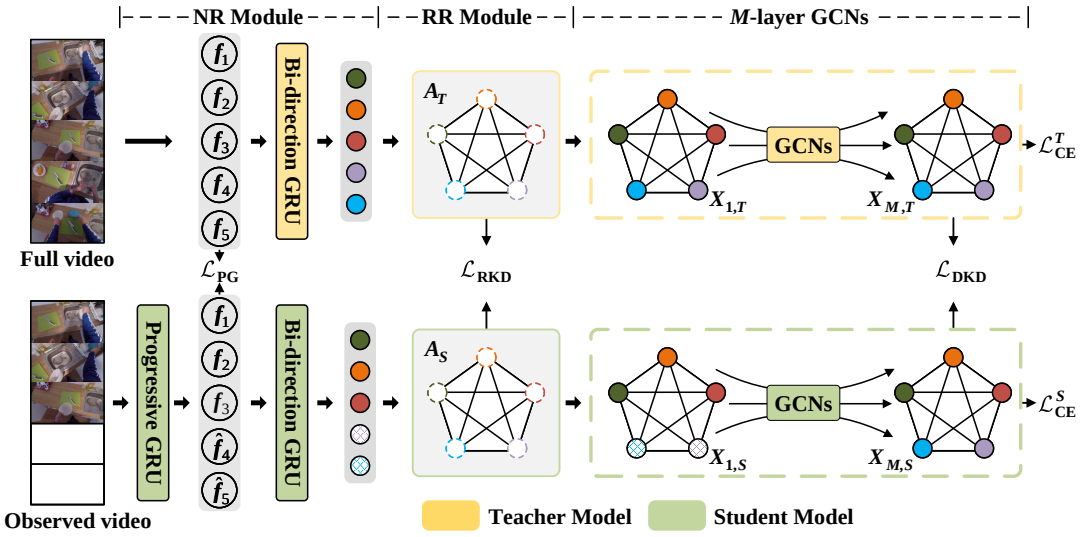

Figure 3: An overview of the proposed Global Relation Graph Network (GRGN). Here we show an example that the observed video has 3 snippets and the unobserved video has 2 snippets. f_i denotes one of the three kinds of modality features, *i.e.*, r_i, m_i or o_i. In student model, a progressive GRU initializes the representations of future video snippets $\{\hat{f}_4, \hat{f}_5\}$ based on f_1, f_2 and f_3. In contrast, the representations of the full video $\{f_1, f_2, \ldots, f_5\}$ are all available for the teacher model. The GRGN, consisting of Node Representation (NR) Module, Relation Representation (RR) Module and a M-layer GCNs, is proposed to reason out discriminative features by modeling global relations between past and future actions. With the knowledge distillation strategy, the privileged relation knowledge learned from the full video can be propagated to the student model.

discriminative feature of the future snippet, which will be further used to predict the action class. Moreover, we adopt a late fusion strategy to combine the predicted results obtained by different kinds of features to exploit the complementarity of different modalities. The teacher model has a similar architecture with the student model, except that the true feature of future video snippet is used to build the relation graph. Under the teacher-student knowledge distillation strategy, the privileged relation knowledge of the past and future actions learned in the teacher model can be propagated to the student model during train phase. At test phase, only the student model is retained to anticipate the future action.

Figure 3 elaborates the GRGN on one of the three kinds of features. More details of the GRGN will be introduced in Section 3.2. The optimization strategy in a teacher-student knowledge distillation manner is introduced in Section 3.3.

3.2 Global Relation Graph Network

Since egocentric video usually changes quickly with the movement of the actor, there exists a clear gap between past and future video snippets making it hard for long-time reasoning in action anticipation task. In our work, the GRGN is designed to infer the feature of the future snippet with the consideration of global relations among observed and unobserved video snippets.

For simplicity, we use f_i to denote the feature of one of the three modalities (*i.e.*, r_i, m_i or o_i) for the observed video snippet. For the unobserved snippet, the feature is defined as $\hat{f}_i$, which is computed in different ways in the teacher model and the student model. After obtaining the representations of all video snippets, the key step of the GRGN is to represent video snippets as a graph structure with discriminative node features and reasonable node relations. We denote the graph by $\mathcal{G}^f = (\mathcal{V}^f, \mathcal{E}^f)$, where $\mathcal{V}^f$ is a set of $l + a$ nodes corresponding to l observed video snippets and a unobserved snippets. The $e^f(i, j) \in \mathcal{E}^f$ denotes the weight of the edge connecting the i-th node and the j-th node. Here, $f \in \{r, m, o\}$ represents the index of the feature modality. The details for building the graph $\mathcal{G}^f$ and reasoning on it are introduced as follows.

3.2.1 Node Representation Module. Here, we introduce how to initialize feature representations $\{\hat{f}_{l+1}, \hat{f}_{l+2}, \ldots, \hat{f}_{l+a}\}$ of unobserved video snippets for the student model by a progressive GRU [4]. For the teacher model, it is much simpler because all graph nodes can be directly represented with the true features. In student model, given the feature f_i of the i-th observed video snippet, the progressive GRU produces the feature $\hat{f}_{i+1}$ of the next unobserved snippet as follows:

$$\begin{aligned} \boldsymbol{h}_{i+1} &= \mathrm{GRU}(\boldsymbol{f}_i, \boldsymbol{h}_i) \\ \hat{\boldsymbol{f}}_{i+1} &= \sigma(\varphi(\boldsymbol{h}_{i+1}) + \boldsymbol{f}_i) \end{aligned} \tag{1}$$

where $\boldsymbol{h}_i$ is the hidden state of GRU, $\varphi(\cdot)$ is a transformation layer and $\sigma(\cdot)$ is a nonlinear activation function ReLU. Here, we use a shortcut connection between $\varphi(\boldsymbol{h}_{i+1})$ and $\boldsymbol{f}_i$ as in [47] to make the progressive GRU aware of the feature difference between two snippets. During anticipation process, we iteratively input the predicted $\hat{\boldsymbol{f}}_i$ into the progressive GRU for initializing the feature $\hat{\boldsymbol{f}}_{i+1}$ of the next snippet.

To effectively predict the correct future action, the temporal information of video snippets are also important because they can provide temporal structure of the past actions. So we input

the features $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_l, \hat{\boldsymbol{f}}_{l+1}, \hat{\boldsymbol{f}}_{l+2}, \ldots, \hat{\boldsymbol{f}}_{l+a}\}$ of all snippets into bi-direction two-layer GRUs to capture more sequential structure features. The encoded forward hidden states in the last GRU layer are represented as $\{\overrightarrow{\boldsymbol{h}}^f_1, \overrightarrow{\boldsymbol{h}}^f_2, \ldots, \overrightarrow{\boldsymbol{h}}^f_{l+a}\}$ and the backward hidden states are represented as $\{\overleftarrow{\boldsymbol{h}}^f_1, \overleftarrow{\boldsymbol{h}}^f_2, \ldots, \overleftarrow{\boldsymbol{h}}^f_{l+a}\}$. The $\overrightarrow{\boldsymbol{h}}^f_i$ or $\overleftarrow{\boldsymbol{h}}^f_i \in \mathbb{R}^{p^f}$ and p^f is the dimension of the hidden states. $f \in \{r, m, o\}$ is the modality index. Then we combine the latent vector representations $\boldsymbol{h}^f_i = [\overrightarrow{\boldsymbol{h}}^f_i; \overleftarrow{\boldsymbol{h}}^f_i] \in \mathbb{R}^{2p^f}$ to obtain representation of each graph node.

3.2.2 Relation Representation Module. We use pairwise similarities of node features to compute the node relations. Specifically, the weight of the edge between the i-th node and the j-th node is defined as follows:

$$e^f(i,j) = \phi^f(\boldsymbol{h}^f_i)^\top \psi^f(\boldsymbol{h}^f_j), \tag{2}$$

where $\phi^f(\cdot)$ and $\psi^f(\cdot)$ represent two transformation functions of node features. The ϕ^f and ψ^f are defined as $\phi^f(\boldsymbol{x}) = \boldsymbol{W}^f_1 \boldsymbol{x}$ and $\psi^f(\boldsymbol{x}) = \boldsymbol{W}^f_2 \boldsymbol{x}$, where $\boldsymbol{W}^f_1$ and $\boldsymbol{W}^f_2$ are trainable parameters with the dimension of $2p^f \times 2p^f$. Subsequently, we use $\boldsymbol{A}^f \in \mathbb{R}^{(l+a)\times(l+a)}$ to denote the adjacent matrix associated to $\mathcal{G}^f$. Each entry of $\boldsymbol{A}^f$ is computed from $e^f(i,j)$ with a row-wise normalization:

$$\boldsymbol{A}^f(i,j) = \frac{\exp\left(e^f(i,j)\right)}{\sum_{j=1}^{l+a} \exp\left(e^f(i,j)\right)}. \tag{3}$$

3.2.3 Reasoning on Graph. For reasoning on the graph, we perform M-layer graph convolutional networks (GCNs) [20] to update the node representations of past video snippets and infer the node representation of future snippet to be anticipated. Specifically, the message passing operation in the m-th GCN layer is conducted as follows:

$$\boldsymbol{X}^f_m = \sigma(\boldsymbol{A}^f \boldsymbol{X}^f_{m-1} \boldsymbol{W}^f_m) + \boldsymbol{X}^f_{m-1} \tag{4}$$

where $\boldsymbol{X}^f_m \in \mathbb{R}^{(l+a)\times d^f_m}$ is the hidden features of all nodes in the m-th GCN layer and d^f_m is the dimension of the hidden feature. $\boldsymbol{X}^f_0 \in \mathbb{R}^{(l+a)\times 2p^f}$ is initialized as $[\boldsymbol{h}^f_1; \boldsymbol{h}^f_2; \ldots; \boldsymbol{h}^f_{l+a}]$. $\boldsymbol{W}^f_m \in \mathbb{R}^{d^f_{m-1}\times d^f_m}$ is a trainable weight matrix and d^f_{m-1} is the dimension of the hidden feature in the $(m-1)$-th GCN layer. For each layer, σ is a nonlinear activation function ReLU. In addition, we use a shortcut connection in each GCN layer to more effectively train the GCN networks and obtain more stable reasoning results.

After the above graph convolution operations, the representation of each node is updated by message passing from neighborhood nodes. More discriminative feature of the future video snippet will also be obtained. To perform action anticipation based on the updated node representations $\boldsymbol{X}^M_f$ of the graph, we apply a fully-connected layer with Softmax activation function:

$$\hat{\boldsymbol{y}}^f = \text{Softmax}(\boldsymbol{W}^f_c \boldsymbol{X}^f_M(l+a) + \boldsymbol{b}^f_c), \quad f \in \{r, m, o\} \tag{5}$$

where $\boldsymbol{W}^f_c$ and $\boldsymbol{b}^f_c$ are trainable parameters. $\boldsymbol{X}^f_M(l+a)$ is the $(l+a)$-th row of $\boldsymbol{X}^f_M$, which denotes the representation of future video snippet.

3.3 Optimization

The optimization of the proposed method has two stages. The teacher model is firstly trained to converge on all training videos, before being used to guide the learning of the student model. Details of the teacher model learning and the student model learning are illustrated as follows.

3.3.1 Teacher Model Learning. The teacher model has a similar architecture with the student model. The only difference of them is that the teacher model directly uses the true features of the future video snippets to build the graph. More specifically, without applying progressive GRU for initializing the node representations, we directly use $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{l+a}\}$ including features of the observed and future video snippets to construct the graph $\mathcal{G}^f$ introduced in Section 3.2. To this end, the action class is much easier to predict since the future video snippets are available and the teacher model is practically an action recognition model based on both observed and unobserved video snippets.

We use Cross Entropy loss to optimize the teacher model:

$$\mathcal{L}^T_{\text{CE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}_T) = -\boldsymbol{y}^\top \log \hat{\boldsymbol{y}}_T, \tag{6}$$

where $\boldsymbol{y}$ is a one-hot vector of the ground-truth, and $\hat{\boldsymbol{y}}_T$ is the probability vector of different action classes predicted by the teacher model. Since our model is based on three kinds of modalities as shown in Figure 2, we first individually train the three GRGNs of different modalities and then synchronously train them based on fused prediction $\hat{\boldsymbol{y}}$ obtained by the late fusion strategy.

3.3.2 Student Model Learning. For the teacher model, since the future video snippets are available, it can conveniently learn the relation knowledge between past and future actions. In this part, our target is to make the student model produce consistent global context relations with the pretrained teacher model, even when the future video snippets are unavailable, so that the privileged relation knowledge between past and future actions learned by the teacher model can be distilled to the student model. Under the constraint of such privileged relation knowledge, the student model will obtain more discriminative features of future video snippets by effectively propagating useful feature information from past video snippets. Specifically, we adopt a cross entropy loss, a progressive GRU loss and two knowledge distillation losses to optimize the student model:

$$\mathcal{L} = \mathcal{L}^S_{\text{CE}}(\bar{\boldsymbol{y}}, \hat{\boldsymbol{y}}_S) + \lambda_0 \mathcal{L}_{\text{PG}} + \lambda_1 \mathcal{L}_{\text{RKD}} + \lambda_2 \mathcal{L}_{\text{DKD}}, \tag{7}$$

where λ_0, λ_1 and λ_2 are hyper-parameters for balancing different losses. $\hat{\boldsymbol{y}}_S$ is the probability vector of different action classes predicted by the student model. $\bar{\boldsymbol{y}}$ is a smoothed action label vector [2] of the groundtruth action class. It is used to not only penalize the error related to the correct action class but also the error related to other similar actions. Here, the similarities between actions are calculated as semantic similarities based on word embeddings of actions obtained by the pretrained word2vector model as in [2]. The remaining items of Eq. (7) are introduced as follows.

Progressive GRU Loss. To optimize the progressive GRU introduced in Section 3.2.1 to efficiently initialize the representations of unobserved video snippets for student model, we adopt a mean

squared error loss between the preliminary feature $\hat{\boldsymbol{f}}_i$ of the unobserved snippet predicted by the progressive GRU and the ground-truth feature $\boldsymbol{f}_i$ as follows:

$$\mathcal{L}_{\mathrm{PG}} = \sum_{f \in \{\boldsymbol{r},\boldsymbol{m},\boldsymbol{o}\}} \sum_{i=l+1}^{l+a} \|\boldsymbol{f}_i - \hat{\boldsymbol{f}}_i\|^2. \tag{8}$$

Relation Knowledge Distillation Loss. We adopt a relation knowledge distillation loss to align the relation matrices of graph nodes computed in the teacher model and the student model. Specifically, for each of the three kinds of features (*i.e.*, $f \in \{r, m, o\}$), we use $\boldsymbol{A}_S^f$, as defined in Eq.(3), to denote the relation matrix calculated in the student model, and use $\boldsymbol{A}_T^f$ to denote the relation matrix calculated in the teacher model. The relation knowledge distillation loss is defined as the KL divergence between these two relation matrices:

$$\mathcal{L}_{\mathrm{RKD}} = -\sum_{f \in \{r,m,o\}} \sum_{i,j=1}^{l+a} A_S^f(i,j) \log\left(\frac{A_T^f(i,j)}{A_S^f(i,j)}\right). \tag{9}$$

Discriminative Knowledge Distillation Loss. To further ensure that the inferred feature of the unobserved video snippet in the student model is discriminative enough for action anticipation, we adopt a discriminative knowledge distillation loss as follows:

$$\mathcal{L}_{\mathrm{DKD}} = \sum_{f \in \{r,m,o\}} \|\boldsymbol{X}_{M,S}^f - \boldsymbol{X}_{M,T}^f\|_F^2, \tag{10}$$

where $\boldsymbol{X}_{M,T}^f$ and $\boldsymbol{X}_{M,S}^f$ are the outputs of M-layer GCNs of teacher and student model. $\|\cdot\|_F$ is the Frobenius norm of matrix.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

EPIC-Kitchens Dataset. This dataset [5] consists of 55 hours of videos collected by 32 participants when they are performing activities in their native kitchen environments. This dataset contains 39,596 video segments with annotated action labels for training and 11,003 segments for testing. There are 125 unique verb classes and 352 unique noun classes in total, while the number of action categories is 2,513. We split the training set of EPIC-Kitchens into training and validation sets by choosing 232 untrimmed videos for training and 40 videos for validation as in [9], resulting in 23,493 segments for training and 4,979 segments for validation.

EGTEA Gaze+ Dataset. The EGTEA Gaze+ [26] dataset contains 10,325 annotated video segments. The annotations have 106 unique action categories. Methods are evaluated on the EGTEA Gaze+ by reporting the average performance of the three splits provided by the authors [26], where each split has 8299 segments for training and 2022 for validation.

Metrics. As in previous works [8, 9, 52], we utilize the metrics of Top-1 Accuracy and Top-5 Accuracy to evaluate our model and compare it with other methods. Moreover, since actions in the EPIC-Kitchens dataset are annotated in the format of (verb, noun) pairs, we also report the anticipation results of predicted verbs and nouns.

4.2 Implementation Details

Similar to the work [9], we employ three kinds of features (*i.e.*, RGB, motion and object features) in the experiments of EPIC-Kitchens dataset. On EGTEA Gaze+ dataset, previous methods [9, 17, 30] only use two kinds of features (*i.e.*, RGB and motion features) since no object annotations are given. We also use these two kinds features as in [9] for fair comparison. The duration τ_o of observed video is set to 2.5 seconds and the time gap τ_α for anticipating is set to 1 second. We segment the video in snippets with the length of $\delta = 0.25$ seconds. Therefore, the total number of observed video snippets is $l = 11$ and the total number of unobserved video snippets is $a = 4$. Here, each video snippet is represented by its first frame. The dimension of the hidden state vector in progressive GRU is set to 1024, 1024 and 352 for the RGB branch, motion branch and object branch, respectively. The dimension of the hidden state vector in bi-GRUs is set to 512, 512 and 176 for three branches. The number of GCN layers in the GRGN is set to 3. The dimension of the node representations is set to 1024 in RGB GRGN, 1024 in motion GRGN and 352 in object GRGN, respectively. In the loss function of the student model illustrated in Eq. (7), the balance weight λ_0, λ_1 and λ_2 are set to 5, 0.1 and 5, respectively . We employ stochastic gradient descent optimizer with a learning rate of 0.01 and momentum of 0.9. The training batch size is 128. All the models are trained for 150 epochs and the model performs best on validation set is used for evaluation on testing set.

Table 1: Anticipation results on S1 test set of EPIC-Kitchens.

Methods	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
	Verb	Noun	Action	Verb	Noun	Action
DMR [42]	26.53	10.43	01.27	73.30	28.86	07.17
2SCNN [5]	29.76	15.15	04.32	76.03	38.56	15.21
ATSN [5]	31.81	16.22	06.00	76.56	42.15	28.21
MCE [7]	27.92	16.09	10.76	73.59	39.32	25.28
ED [11]	29.35	16.07	08.08	74.49	38.83	18.19
Miech *et al.* [32]	30.74	16.47	09.74	76.21	42.72	25.44
RULSTM [9]	33.04	22.78	14.39	79.55	50.95	33.73
Liu *et al.* [27]	34.99	20.86	14.04	77.05	46.45	31.29
SRL [36]	34.89	22.84	14.24	79.59	52.03	34.61
KDLM [2]	35.04	23.03	14.43	79.56	52.90	34.99
ImagineRnn [47]	35.44	22.79	14.66	**79.72**	52.09	34.98
Sener *et al.* [38]	37.90	24.10	16.60	79.70	**54.00**	36.10
MGRKD	**38.70**	**25.20**	**16.98**	79.15	53.44	**37.12**

4.3 Comparison with State-of-the-Arts

EPIC-Kitchens. On this dataset, we compare our model with following competitive methods: DMR [42], 2SCNN [5], ATSN [5], MCE [7], ED [11], Miech *et al.* [32], RULSTM [9], RU+IAI [52], Liu *et al.* [27], SRL [36], KDLM [2] and ImagineRnn [47]. As in existing works [8, 9], we report the results on both the seen test set (S1), where the test videos may have scenes appear in the training videos, and the unseen test (S2), where the test videos do not have the same scenes with the training videos. It is worth noting that Liu *et al.* [27] use interaction hotspots and hand trajectory features in action anticipation. Since most of existing methods use RGB, motion and object features, we report the results without using hotspots and hand trajectory in [27] for fair comparison. As show in Table 1, our MGRKD outperforms all other methods by a significant margin on the S1 set. Moreover, on the S2 set as shown in Table 2, our MGRKD also achieves competitive results. The results

Table 2: Anticipation results on S2 test set of EPIC-Kitchens.

Methods	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
	Verb	Noun	Action	Verb	Noun	Action
DMR [42]	24.79	08.12	00.55	64.76	20.19	04.39
2SCNN [5]	25.23	09.97	02.29	68.66	27.38	09.35
ATSN [5]	25.30	10.41	02.39	68.32	29.50	06.63
MCE [7]	21.27	09.90	05.57	63.33	25.50	15.71
ED [11]	22.52	07.81	02.65	62.65	21.42	07.57
Miech *et al.* [32]	28.37	12.43	07.24	69.96	32.20	19.29
RULSTM [9]	27.01	15.19	08.16	69.55	34.38	21.10
RU-IAI [52]	27.89	14.89	08.57	70.06	35.51	21.41
Liu *et al.* [27]	28.27	14.07	08.64	70.67	34.35	22.91
SRL [36]	27.42	15.47	08.88	**71.90**	36.80	22.06
KDLM [2]	29.29	15.33	08.81	70.71	36.63	21.34
ImagineRnn [47]	29.33	15.50	09.25	70.67	35.78	22.19
Sener *et al.* [38]	**29.50**	16.50	10.10	70.10	**37.80**	**23.40**
MGRKD	29.26	**16.59**	**10.38**	70.78	37.32	23.05

Table 3: Top-5 accuracy (%) results on EGTEA Gaze+ with different values of the anticipation time gap.

Methods	Top-5 Action Accuracy @ Anticipation Time τ_α							
	2.0	1.75	1.5	1.25	1.0	0.75	0.5	0.25
DMR [42]	-	-	-	-	55.70	-	-	-
ATSN [5]	-	-	-	-	40.53	-	-	-
MCE [7]	-	-	-	-	56.29	-	-	-
ED [11]	45.03	46.22	46.86	48.36	50.22	51.86	49.99	49.17
FN [6]	54.06	54.94	56.75	58.34	60.12	62.03	63.96	66.45
RL [30]	55.18	56.31	58.22	60.35	62.56	64.65	67.35	70.42
EL [17]	55.62	57.56	59.77	61.58	64.62	66.89	69.60	72.38
rulstm [9]	56.82	59.13	61.42	63.53	66.40	68.41	71.84	74.28
ImagineRnn [47]	-	-	-	-	66.71	68.54	72.32	74.59
KDLM [2]	59.99	62.02	63.95	66.47	68.74	72.16	75.21	78.11
SRL [36]	59.69	61.79	64.93	66.45	70.67	73.49	**78.02**	**82.61**
MGRKD	**60.86**	**63.43**	**65.24**	**67.66**	**70.86**	**74.32**	77.49	79.61

of our method on action anticipation are better than other models. For Verb anticipation, though the proposed MGRKD can not outperform the ImagineRnn [47] on the metrics of Top-1 and Top-5 accuracy, we consistently achieve competitive results among all the compared methods on all metrics. The comparison results in Table 1 and 2 demonstrate the effectiveness of proposed model.

EGTEA Gaze+. As in existing works [9], we report Top-5 accuracy results of action anticipation for comparison. Moreover, we also evaluate our model with different anticipation time gaps (*i.e*, τ_α is set to $\{2.0, 1.75, \ldots, 0.25\}$). As shown in Table 3, the proposed MGRKD outperforms all the other methods on all time gaps, except when $\tau_\alpha = \{0.5, 0.25\}$. It is worth noting that our MGRKD consistently achieves better performances than SRL when τ_α is larger than 0.5, which demonstrates that our model is more effective in capturing the dependencies with long time intervals for action anticipation.

4.4 Ablation Studies

We conduct ablation studies in this part to further illustrate the effectiveness of our method. The results on EPIC-Kitchens dataset are reported on validation set and the results on EGTEA Gaze+ dataset are reported as the average performance of three splits.

Knowledge Distillation Strategy. To evaluate the effectiveness of the knowledge distillation strategy in our MGRKD, we consider several variants of our model learned with different loss functions. Specifically, as shown in Table 4, the Student w/o KD denotes the student model trained without both the Relation Knowledge Distillation loss and the Discriminative Knowledge Distillation loss illustrated in Eq. (9) and (10). The Student w/o DKD denotes the model trained without the Discriminative Knowledge Distillation loss. The Student w/o RKD denotes the model trained without the Relation Knowledge Distillation loss. The Teacher model uses true features of future video snippets for action recognition, which performs as an upper-bound of all student models. The experimental results show that the student model without knowledge distillation learning strategy can not perform well in action anticipation. Moreover, the RKD loss is very effective to improve the student model. Although the DKD loss can distill discriminative features from the teacher model to the student model, it achieves less improvements for action anticipation compared with RKD. These results further demonstrate the effectiveness of the proposed MGRKD. It is worth noting that, comparing student model with student w/o KD, the improvements on EGTEA Gaze+ dataset are not as significant as those on EPIC-Kitchens (*i.e.*, +1.62% on EGTEA Gaze+ versus +3.01% on EPIC-Kitchens with respect to Top-5 Accuracy). Because EGTEA Gaze+ has fewer action categories compared with EPIC-Kitchens (106 actions versus 2,513 actions), which always results in fewer actions (*e.g.*, 1 or 2 actions) in the observed videos. Therefore, exploring global relations between past and future actions on EGTEA Gaze+ has less impact on action anticipation.

Table 4: Ablation studies of distillation strategy on two datasets. The Top-1/5 Accuracies (Acc@1/5) of action anticipation with $\tau_\alpha = 1$ are reported.

Methods	EPIC-Kitchens		EGTEA Gaze+	
	Acc@1	Acc@5	Acc@1	Acc@5
Teacher	23.23	44.97	49.98	82.52
Student w/o KD	15.52	34.98	36.69	69.24
Student w/o DKD	16.81	36.38	36.83	69.88
Student w/o RKD	15.97	35.79	36.78	69.68
Student (MGRKD)	**17.22**	**37.99**	**37.80**	**70.86**

Table 5: Ablation studies of modality fusion on two datasets. The Top-1/5 Accuracies (Acc@1/5) of action anticipation with $\tau_\alpha = 1$ are reported.

Methods	EPIC-Kitchens		EGTEA Gaze+	
	Acc@1	Acc@5	Acc@1	Acc@5
RGB	14.88	33.42	36.91	69.89
Motion	09.53	22.69	15.78	43.34
Object	11.93	31.42	-	-
MGRKD	**17.22**	**37.99**	**37.80**	**70.86**

Modality Fusion Strategy. Table 5 shows ablation studies on the modality fusion strategy. We evaluate the performances of our model using only one kind of features, *i.e.*, RGB, motion or object. As shown, results of our full model are significantly better than the ones using only a single modality. We can conclude that considering the complementarity among different modalities is very important.

Table 6: Ablation studies of node representation module on two datasets. The Top-1/5 Accuracies (Acc@1/5) of action anticipation with $\tau_\alpha = 1$ are reported.

Methods	EPIC-Kitchens		EGTEA Gaze+	
	Acc@1	Acc@5	Acc@1	Acc@5
MGRKD w/o PG	15.47	34.88	35.18	67.59
MGRKD w/o Bi-GRUS	14.84	34.86	32.96	65.13
MGRKD	**17.22**	**37.99**	**37.80**	**70.86**

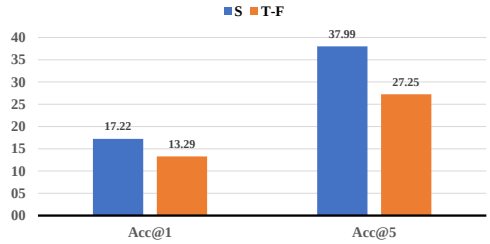

Figure 4: Comparison of the teacher model (T-F) and the student model (S) in action anticipation on EPIC-Kitchens.

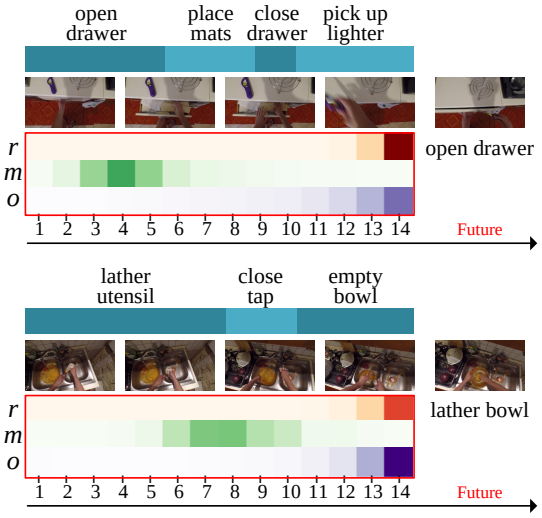

Figure 5: Qualitative example of past actions that have distinguished edges with the future action. r denotes the GRGN with RGB features, m denotes the GRGN with motion features, o denotes the GRGN with object features.

Node Representation Analysis. Table 6 shows ablation studies on the node representation module illustrated in Section 3.2.1. We evaluate two variants of the proposed node representation, MGRKD w/o PG and MGRKD w/o Bi-GRUs. The MGRKD w/o PG denotes the model simply using zero-vectors as the initialized node representations of future video snippets. The MGRKD w/o Bi-GRUs denotes the model removing the Bi-direction GRUs introduced in Section 3.2.1. As shown, MGRKD performs significantly better than MGRKD w/o PG and MGRKD w/o GRU, which demonstrates the effectiveness of the proposed node representation scheme.

4.5 Further Remarks

Comparison between Teacher and Student Model. As illustrated in Section 3.3, the student model solves the action anticipation task, while the teacher model practically solves an easier task of action recognition. In our MGRKD, the major role of the teacher model is to provide the privileged relation knowledge between the past and future actions for the student model. Therefore, if the teacher model can provide extremely useful knowledge to solve the action anticipation task, an interesting question is raised: whether the teacher model can tackle the action anticipation task by itself ? To answer this question, we directly apply the teacher model (it has been well trained on the action recognition task) to the action anticipation task by replacing the node feature of the future video snippet in the GRGN with the feature initialized by the progressive GRU. We denote this variant of the teacher model as T-F. As shown in Figure 4, the variant teacher model (T-F) performs worse than the student model (S), which demonstrates the necessity of the designed teacher-student learning strategy in our MGRKD.

Qualitative Results. In our model, the global relation graph networks (GRGN) illustrated in Section 3.2 are built to infer the discriminative feature of future video by globally considering pairwise relations between past and future actions. Figure 5 shows some examples of past actions that have distinguished edges (the edges with large weights in the GRGN) with the future action. Specifically, taking the top part of Figure 5 as an example, our model focuses on "*pick up lighter*" in the GRGN with RGB features and the GRGN with object features. In the GRGN with motion features, the action "*open drawer*" provides more important information to anticipate the future action "*open drawer*". This relation knowledge is important to anticipate the correct action.

5 CONCLUSION

In this paper, we propose a Multimodal Global Relation Knowledge Distillation (MGRKD) framework which implement a teacher-student learning strategy for action anticipation. Either the teacher or student model consists of three branches of global relation graph networks (GRGN) which can explore the pairwise relations between the past and future actions based on three kinds of features (*i.e.*, RGB, motion or object). The student model builds a relation graph based on the observed video snippets to reason out discriminative feature of future snippet, which will be further used to predict action class. The predicted results of different GRGNs are fused together to improve the performance. The teacher model has a similar architecture with student model, except that the true features of the unobserved video snippets are used to build the relation graph. Extensive experimental results demonstrate the effectiveness of the proposed method.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (No. 2018AAA0100604), National Natural Science Foundation of China (No. 61720106006, 62036012, 61721004, 62072455, U1836220, U1705262), Key Research Program of Frontier Sciences of CAS (QYZDJ-SSW-JSC039), Beijing Natural Science Foundation (L201001), CASIA-LLVision Joint Lab.

REFERENCES

[1] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. 2018. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4194–4202.

[2] Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni Maria Farinella, and Lamberto Ballan. 2020. Knowledge distillation for action anticipation via label smoothing. In *25th International Conference on Pattern Recognition*.

[3] Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, and Zhe Ma. 2021. ECKPN: Explicit Class Knowledge Propagation Network for Transductive Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision*. 720–736.

[6] Roeland De Geest and Tinne Tuytelaars. 2018. Modeling temporal structure with LSTM for online action detection. In *2018 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1549–1557.

[7] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. 2018. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *European Conference on Computer Vision*.

[8] Antonino Furnari and Giovanni Farinella. 2020. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[9] Antonino Furnari and Giovanni Maria Farinella. 2019. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 6252–6261.

[10] Junyu Gao, Xiaoshan Yang, Yingying Zhang, and Changsheng Xu. 2020. Unsupervised video summarization via relation-aware assignment learning. *IEEE Transactions on Multimedia* (2020).

[11] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. 2017. Red: Reinforced encoder-decoder networks for action anticipation. In *British Machine Vision Conference*.

[12] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8303–8311.

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2827–2836.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[15] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang. 2018. Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (2018), 2568–2583.

[16] Yifei Huang, Yusuke Sugano, and Yoichi Sato. 2020. Improving Action Segmentation via Graph-Based Temporal Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 14024–14034.

[17] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. 2016. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation*. IEEE, 3118–3125.

[18] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. 2019. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11487–11496.

[19] Takeo Kanade and Martial Hebert. 2012. First-person vision. *IEEE* 100, 8 (2012), 2442–2453.

[20] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

[21] Yu Kong, Zhiqiang Tao, and Yun Fu. 2017. Deep sequential context networks for action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1473–1481.

[22] Hema S Koppula and Ashutosh Saxena. 2015. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 1 (2015), 14–29.

[23] Anoop Korattikara Balan, Vivek Rathod, Kevin P Murphy, and Max Welling. 2015. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems*, Vol. 28. 3438–3446.

[24] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *European Conference on Computer Vision*. Springer, 689–704.

[25] Kang Li, Jie Hu, and Yun Fu. 2012. Modeling complex temporal composition of actionlets for activity prediction. In *European Conference on Computer Vision*. Springer, 286–299.

[26] Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision*. 619–635.

[27] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. 2020. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*. Springer, 704–721.

[28] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2015. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643* (2015).

[29] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, Xiaoou Tang, et al. 2016. Face Model Compression by Distilling Knowledge from Neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3560–3566.

[30] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1942–1950.

[31] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. 2019. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6120–6127.

[32] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. 2019. Leveraging the present to anticipate the future in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[33] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*. 2654–2665.

[34] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. 2020. Spatio-Temporal Graph for Video Captioning with Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10870–10879.

[35] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2020. Emotion knowledge driven video highlight detection. *IEEE Transactions on Multimedia* (2020).

[36] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. 2021. Self-Regulated Learning for Egocentric Video Activity Anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[37] MS Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. 2015. Robot-centric activity prediction from first-person videos: What will they do to me?. In *10th International Conference on Human-Robot Interaction*. IEEE, 295–302.

[38] Fadime Sener, Dipika Singhania, and Angela Yao. 2020. Temporal Aggregate Representations for Long Term Video Understanding. In *European Conference on Computer Vision*.

[39] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568–576.

[40] Bilge Soran, Ali Farhadi, and Linda Shapiro. 2015. Generating notifications for missing actions: Don't forget to turn the lights off!. In *Proceedings of the IEEE International Conference on Computer Vision*. 4669–4677.

[41] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 98–106.

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

[44] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. 2020. Learning coarse-to-fine graph neural networks for video-text retrieval. *IEEE Transactions on Multimedia* (2020).

[45] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. 2019. Progressive teacher-student learning for early action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3556–3565.

[46] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6857–6866.

[47] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. 2021. Learning to Anticipate Egocentric Actions by Imagination. *IEEE Transactions on Image Processing* 30 (2021), 1143–1152.

[48] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. 2020. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3289–3299.

[49] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *European Conference on Computer Vision*. 684–699.

[50] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 7094–7103.

[51] Jingran Zhang, Fumin Shen, Xing Xu, and Heng Tao Shen. 2020. Temporal reasoning graph for activity recognition. *IEEE Transactions on Image Processing* 29 (2020), 5491–5506.

[52] Tianyu Zhang, Weiqing Min, Ying Zhu, Yong Rui, and Shuqiang Jiang. 2020. An Egocentric Action Anticipation Framework via Fusing Intuition and Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 402–410.

[53] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. 2019. A structured model for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9975–9984.

[54] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More Grounded Image Captioning by Distilling Image-Text Matching Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4777–4786.