HARDer-Net: Hardness-Guided Discrimination Network for 3D Early Activity Prediction

Tianjiao Li, Yang Luo, Wei Zhang, Lingyu Duan, and Jun Liu

***Abstract*—To predict the class label from a partially observable activity sequence can be quite challenging due to the high degree of similarity existing in early segments of different activities. In this paper, an innovative HARDness-Guided Discrimination Network (HARDer-Net) is proposed to evaluate the relationship between similar activity pairs that are extremely hard to discriminate. To train our HARDer-Net, an innovative adversarial learning scheme has been designed, providing our network with the strength to extract subtle discrimination information for the prediction of 3D early activities. Moreover, to enhance the adversarial learning scheme efficacy of our model for 3D early action prediction, we construct a Hardness-Guided bank that dynamically records the hard similar samples and conducts reward-guided selections of these recorded hard samples using a deep reinforcement learning scheme. The proposed method significantly enhances the capability of the model to discern fine-grained differences in early activity sequences. Several widely-used activity datasets are used to evaluate our proposed HARDer-Net, and we achieve state-of-the-art performance across all the evaluated datasets.**

***Index Terms*—Early Activity Prediction, 3D Skeleton Data, Action/Gesture Understanding, Hardness-Guided Learning.**

I. Introduction

As an important and prevalent research topic in the field of human behavior understanding, early activity prediction focuses on predicting the class label before action is entirely performed, and it has many real-world applications including online interactions between humans and robots, autonomous vehicles, and surveillance systems [1]–[3]. Existing studies [4]–[12] indicate that 3D skeletal structure data, readily obtainable from low-cost depth cameras, provides a concise yet effective representation of human behaviors. Therefore, the primary objective of this paper is to accurately predict the action categories before human activities are fully executed given 3D skeleton data, which is also known as 3D early activity prediction.

In the context of 3D early activity prediction, observation is confined to the initial parts of the sequences, instead of the entire skeleton sequence (as in 3D activity recognition) which contains adequate discrimination information. Consequently, predicting human activities in very early stages is a much more challenging task compared to a typical action recognition task. More specifically, in the context of the early prediction of human activities, the beginning segments observed in many activities can be very similar, which merely contain minor differences, which are hard for the prediction models to perceive.

Therefore, these partially observable segments containing inadequate discrimination information can be easily miscategorized. For instance, as shown in Fig. 1, the action "pointing to someone" can be wrongly classified into the action "shaking hand" with only slight differences at the early stage (e.g., 20% observation ratio). We refer to segments that are prone to misprediction as *hard instances*, and *interference classes* are classes that *hard instances* are readily mispredicted into. Similarly, a pair consisting of a *hard instance* and the *interference class* is termed a *hard pair*.

In order to address the challenge of the 3D early activity prediction problem, many researchers [2], [13] attempt to distill the global information of the full sequence of activity, which possesses additional information on discrimination, in order to aid in the prediction of activity from the partial sequence of activity, which contains less discriminative information. Although the previous approaches [1], [3], [14] have made remarkable progress, most of these works do not explicitly address the issue of discrimination for *hard pairs*, which is to identify and exploit the slight yet significant discrepancies within each *hard pair* to improve early activity prediction performance.

As we have already highlighted, the subtle differences between the partial observations of the *hard instance* and the corresponding *interference class* give rise to higher "hardness" of 3D early activity prediction task. Therefore, to ensure accurate predictions of partially observed human actions, a recognition model should be capable of grasping the relationship existing in confusing *hard pair* samples and scrutinizing the inherent subtle differences that can be adopted for discrimination.

In light of this, we develop a discriminative model to explicitly exploit the intrinsic discrimination information between the *hardest instance* and its corresponding *interference class*, namely Harderness-Guided Discrimination Network (HARDer-Net), for 3D early activity prediction. To be more specific, as part of our HARDer-Net, a Hardness-Guided bank (HG bank) is developed to be capable of adaptively recording and sampling the *hard pairs* during the model learning procedure. Notably, the proposed HG bank is a

Tianjiao Li is with ISTD Pillar, Singapore University of Technology and Design, and School of Control Science and Engineering, Shandong University. (email: tianjiao_li@mymail.sutd.edu.sg)

Yang Luo is with School of Computing, National University of Singapore. (email: yangluo@comp.nus.edu.sg)

Wei Zhang is with School of Control Science and Engineering, Shandong University. (email: davidzhang@sdu.edu.cn)

Lingyu Duan is with School of EE and CS, Peking University. (email: lingyu@pku.edu.cn)

Jun Liu is with School of Computing and Communications, Lancaster University, UK, and ISTD Pillar, Singapore University of Technology and Design. (email: j.liu81@lancaster.ac.uk)

⋆ Tianjiao Li and Yang Luo contribute equally.

✉ Corresponding author: Jun Liu and Wei Zhang

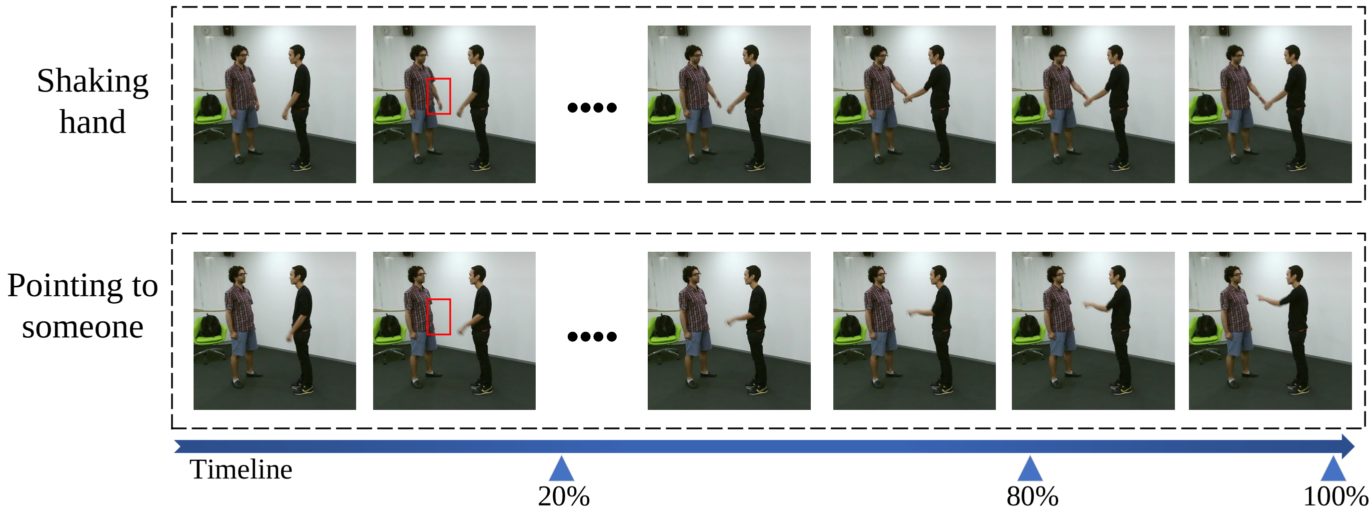


Fig. 1. The figure above illustrates two examples of activities taken from the NTU RGB+D dataset [4]. It is possible to easily differentiate these two activities using adequate discrimination information if their complete sequences are observed, however when these two activities are observed at an early stage (e.g., only 20% of the sequence is observed), they are almost the same (showing only subtle discrimination information, as indicated in the red boxes).

reward-driven reply memory. This means that our newly designed HG bank provides the most informative hard pairs that can directly boost prediction performances. Concretely, the selection of hard pairs is transformed into a decision-making process. We utilize the deep Q-Network (DQN) algorithm [15], [16] and set the prediction accuracy as final reward, which enables our prediction model to focus on maximizing the action prediction performances. Unlike random selection that deployed in our previous conference paper [17], which makes uninformed choices without any strategic consideration, our newly-designed HG bank can continuously evaluate and update the value of actions based on their potential to obtain higher prediction accuracy. This reward-centric approach ensures that each selection of hard pairs is optimized to enhance ultimate rewards, leading to better action prediction performances. By focusing on those hard pairs that contain subtle yet significant cues, the proposed HG bank effectively boosts overall efficiency and effectiveness.

Then the selected representative hard pairs are sent to a feature generator, which is designed to explore the relationship between a *hard instance* and its corresponding *interference class*. The proposed feature generator is able to produce perplexing but conceivable features for the *hard instance* conditioned on its similarities to the corresponding *interference class*. Followed by the feature generator, a class discriminator is introduced to empower the prediction model with the ability to discriminate the perplexing features of the *hard instance* from its corresponding *interference classes*. Accordingly, the generated features get increasingly confusing when viewed from the perspective of its *interference class* as the adversarial learning process going, thereby enhancing the ability of the class discriminator in exploiting the minor discrepancies inside of the features of *hard pair* samples for class discrimination. Consequently, the proposed HARDer-Net with its class discriminator as the classifier is highly effective at dealing with *hard pairs* that are usually remarkably challenging to distinguish using existing early activity prediction models.

The major contributions of this paper can be summarized as follows:

- We propose the Hardness-Guided Discrimination Network, namely HARDer-Net, to alleviate the high similarity issue by explicitly mining the subtle differences between the *hard instance* and its corresponding *interference class* via an adversarial learning scheme.
- A Hardness-Guided bank is also introduced to record the *hard pairs* on the fly and to adaptively select the most representative pairs by using a reward-driven DRL strategy, to directly promote the action prediction performances.
- The proposed HARDer-Net achieves promising performances across four challenging datasets for 3D early activity prediction, which demonstrates the efficacy of our method.

This paper is an extension of our previous conference paper [17]. We clarify the innovations and improvements in the following aspects: **(1) Improved RL-based Hard Sample Mining:** In our original submission, we introduced two major innovations which are (i.) a HI-IC bank mechanism to store hard example pairs, and (ii.) an adversarial learning scheme to exploit the subtle differences between these pairs. However, in our original submission, the hard pairs are randomly selected from the HI-IC bank, which may ignore the most representative pairs in the bank. To address this issue, in this submission, we introduce the upgraded Hardness-Guided (HG) bank which employs a deep reinforcement learning (DRL) scheme to guide the training process directly with the rewards. Compared to the original randomly-sampled HI-IC bank, the updated HG bank is able to select the most informative and representative hard pairs; **(2) New Theoretical Insights:** In this submission, our newly-designed HG bank is reward-driven replay memory and the reward is provided by the ultimate goal, i.e., the recognition performance. Therefore, encouraged by the reward, the HG bank is able to provide the most representative hard pairs containing informative subtle cues that can directly boost the action prediction performances; **(3) More Comprehensive Evaluations:** In the original submission, we provided detailed experimental analyses on NTU RGB+D and FPHA. However, in this submission, to further comprehensively evaluate the efficacy of our method. We extended two additional datasets, i.e., SYSU 3D HOI and UCF101, which are widely accepted in early action prediction. As shown in the experiment section in this submission, our HARDer-Net outperforms other state-of-the-art approaches significantly. Also, we have conducted extensive ablation experiments on our newly designed HG bank across all four datasets. The results demonstrate that the DRL-based reward-driven HG bank can help to exploit more informative subtle cues to benefit the ultimate early action prediction performances.

Below is a summary of this paper. In section II, we discuss the related works. In section III, we describe in greater detail our proposed HARDer-Net for 3D early action prediction. In section IV, we provide the experimental results and comprehensive analyses. At the end, we present the conclusion in section VI.

II. Related Work

Human Activity Recognition. Human activity recognition, a focal point of interest within the deep learning community, attracts the attention of numerous researchers and stands as a prevalent research topic. There exist many approaches [4], [18]–[22] for recognizing 3D human activity utilizing RNN and LSTM-based architectures. Besides, convolutional neural networks (CNNs) [23]–[25] and self-attention networks [26] are also developed for human action recognition. Recently, graph convolutional networks (GCNs) have gained increasing popularity in recent years and been investigated in representing human actions [27]–[32] due to the powerful representative capabilities. Yan *et al.* [28] presented the utilization of spatial-temporal Graph Convolutional Networks (GCN) for the purpose of addressing 3D human activity recognition tasks. Shi *et al.* [27] introduced an adaptive GCN that flexibly learns the topology of each layer in the graph and advances performance by adding second-order data from the original skeleton data as an additional input stream. Besides, Chen *et al.* [33] proposed a hierarchical pyramid structure designed to effectively model multi-scale spatio-temporal information and integrate action information of various granularities.

Early Human Activity Prediction. As opposed to full-length human activity recognition, which can access the whole activity sequences containing abundant discrimination information, early human activity prediction can only observe partial segments of activity sequences from the beginning. This inherent limitation renders the early activity prediction task considerably more challenging in comparison to the typical activity recognition task. There have been several approaches. Most of the existing approaches [1]–[3], [13], [34]–[43] focus on alleviating the information gaps better the full-length and partial-length activity sequences. Ke *et al.* [13] introduced to rely on partial activity sequences for gaining latent local information and full activity sequences for gaining latent global information. Wang *et al.* [2] proposed a method for transferring knowledge from long-term to shorter-term activity sequences through a teacher-student learning architecture. Guan *et al.* [43] constructed transformer-based model by adopting two transformer encoders for extracting features of observed and unobserved actions respectively. Zheng *et al.* [44] introduced an adversarial knowledge distillation (AKD) to transfer the knowledge from a teacher network (optimized by full videos) to a student network (optimized by partial videos). Then a discriminator is employed to encourage the features produced by the student network to approach the features learned from full videos by the teacher network, to enhance the latent representations. However, in our HARDer-Net, we propose to record those hard samples that may cause ambiguities in an HG bank and search for the most informative pairs for our adversarial learning scheme via deep reinforcement learning. Besides, in our adversarial learning scheme, we aim to generate ambiguous latent features to boost the recognition abilities in distinguishing subtle cues for our prediction model.

Nevertheless, the existing works mentioned above do not primarily focus on promoting the discrimination capability of the prediction model by exploiting the extremely similar *hard pair* samples, which is considered to be a limitation of early activity prediction. In contrast to these works, we establish an HG bank to memorize the *hard pair* samples explicitly and dynamically and propose an innovative HARDer-Net conditioned on adversarial learning, which enables our prediction model to discriminate *hard pair* samples by comprehending the relationships between them.

Hard Example Learning. It is widely recognized that explicitly learning from hard examples could be beneficial to the model learning process [45]–[52]. To be more specific, Shrivastava *et al.* [46] introduced an example mining system that autonomously selects challenging data in order to enhance the performance of object classification. Felzenszwalb *et al.* [51] proposed an iterative procedure for fixing the latent values for positive examples and optimizing the objective function of the latent SVM for the handling of hard negative examples using a margin-sensitive SVM.

Different from the aforementioned approaches that focus on learning certain hard examples, we concentrate on improving the capacity to analyze slight discrimination information inside of *hard pairs*, which is comprised of a *hard instance* and its relevant *interference class*. Here note that we utilize the adversarial learning scheme to pair the mispredicted activity segments with their *interference classes*. In addition, an HG bank is further created to memorize the *hard pairs*, aiming to iteratively expedite comprehension of relationships and subtle differences within the pairs. In this way, the early activity prediction model becomes more discriminative.

Reinforcement Learning. Reinforcement learning (RL) [53], [54] focuses on maximizing the cumulative rewards by training a decision maker (i.e., an agent) to take consecutive actions in a prescribed environment. To map the states from a high-dimensional space to a relatively low-dimensional space, deep reinforcement learning (DRL) is further proposed to combine deep neural networks and traditional reinforcement learning algorithms together, i.e., representing the decision-making process using deep neural networks. For instance,

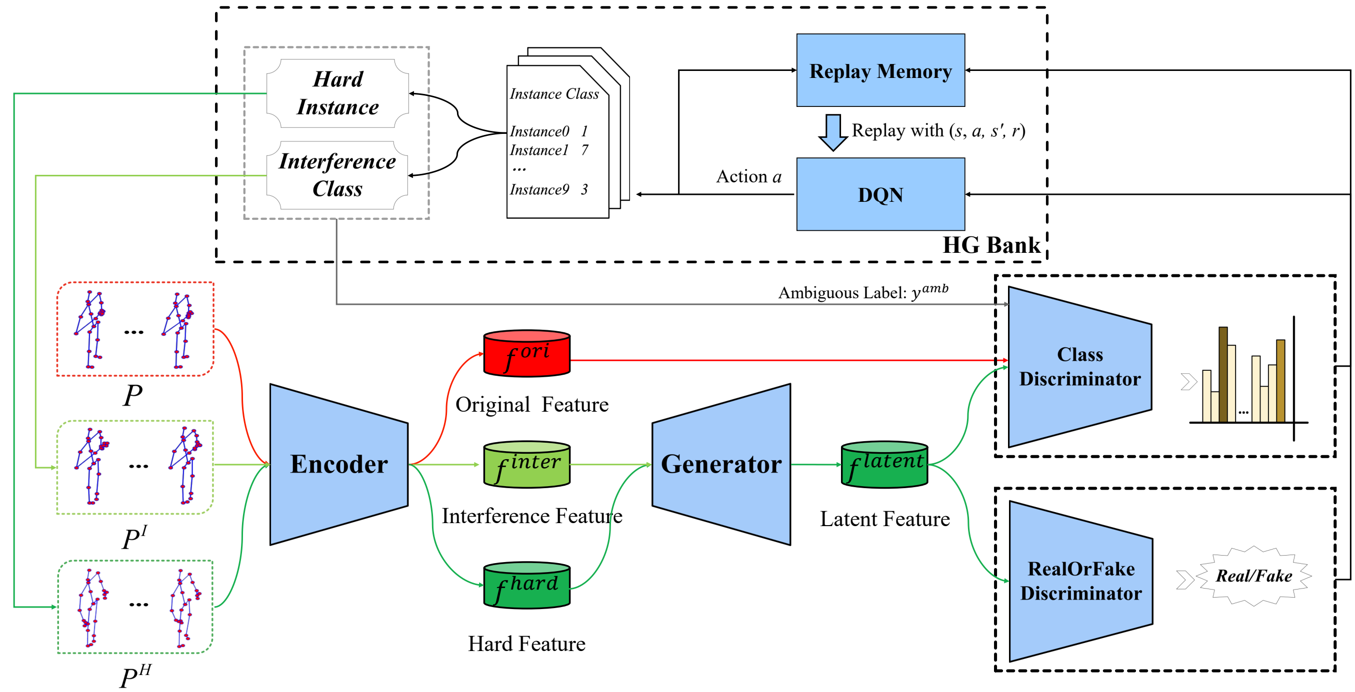


Fig. 2. Overall framework of our end-to-end HARDer-Net. It is established using a substitutable feature encoder (such as a CNN [13] or GCN skeleton encoder [27] which encodes partial sequence features). As indicated by the red arrows in both the training and inference phases, partial activity sequences (P) are transmitted to the encoder, and then to the classifier, for the purpose of obtaining classification scores determining the criteria for storing *hard pairs* in our HG bank. As shown by the green arrows indicating the adversarial learning phase, the HG bank adaptively provides a *hard pair* which contains a *hard instance* and an *interference class* sample for feature encoding, resulting in an increased capability for our prediction model to identify minor differences within each *hard pair* by employing adversarial learning. The grey arrow presents the introduction of "ambiguous label" based on *hard instance* and its *interference class*. With respect to our HG Bank, the black arrows indicate that the state s and reward r are transferred from our two discriminators to the DQN to generate the action a for sampling the *hard pairs*. As for the training phase of the deep reinforcement learning network part, the DQN is trained using tuples (s, a, s', r) that have been saved in replay memory.

DQN [16], which is cast as an extension of the traditional Q-learning algorithm [55], applies deep neural networks in order to approximate the action-value function. The effectiveness of RL-based approaches has been substantially demonstrated in various computer vision domains, including video object segmentation [56] and Vision-and-language Navigation [57].

III. Method

A. Problem Formulation

Considering a full-length action sequence $S = \{s_t\}_{t=1}^{T}$, where s_t represents the t_{th} frame, and T denotes the length of the action sequence. We follow the previous works [1], [13] and divide the full action sequence S into N segments, each of which then contains $\frac{T}{N}$ action frames. Thus a partial sequence is denoted as $P = \{s_t\}_{t=1}^{\tau}$, where $\tau = i \cdot \frac{T}{N}$ and $i \leq N$. Here we define the observation ratio $r_i = \frac{i}{N}$ And the early action prediction task aims to identify the action class $c \in \mathbb{C} = \{1, 2, ..., C\}$ to which the partial activity sequence P corresponds, under varying observation ratios.

B. Hardness-Guided Discrimination Network

1) Overview: As depicted in Fig. 2, our proposed Hardness-Guided Discrimination Network (HARDer-Net) consists of two primary components, the *Adversarial Hardness-Guided Discrimination Learning Scheme* and the *Hardness-Guided Bank (HG Bank)*. As aforementioned, certain activities can exhibit notable similarities during their initial stages. Therefore, the 3D early activity prediction performances are prone to the deficiency of adequate discrimination information, particularly in scenarios where observation ratios are low. By designing an HG bank, we present an innovative method to explicitly memorize the *hard pairs* which are vulnerable to insufficient discrimination information. In the meantime, we introduce the Adversarial Hardness-Guided Learning scheme to explore the correlation between each *hard pair*, i.e., the *hard instance* and the relevant *interference class*. Through iteratively memorizing and exploiting the *hard pairs*, our prediction model can extract subtle yet discriminative information within the feature space to enhance the accuracy of recognition.

2) Adversarial Hardness-Guided Discrimination Learning Scheme: With the aim of generating confusing yet plausible features based on the *hard pairs*, a feature generator is introduced to harness the association between the *hard instance* and their respective *interference class*. Besides, to boost the capability of our network to extract minor discrimination information, a class discriminator ($\mathbb{D}^{cls}$) is further constructed to differentiate between the synthesized latent features regarding the *hard instance* and the corresponding *interference class*.

Through the introduced adversarial hardness-guided learning scheme, the synthesized latent features of the *hard instance* become increasingly perplexing regarding its *inter-*

ference class, resulting in an increased ability for the class discriminator to utilize the slight discrimination information to differentiate between the confounding *hard instance* and the corresponding *interference class*. In this way, adversarial learning strengthens the effectiveness of the prediction model for discrimination.

Feature Generator. We aim to construct a feature generator ($\mathbb{G}$) that leverages the association between the *hard instance* and its *interference class*, thereby creating latent features that are challenging and obscure to predict, but at the same time preserve inherent information associated with the original *hard instance*.

To be more specific, for each hard instance (P^H) in our established HG bank, we adaptively sample an interference instance (P^I) from the *interference class* accordingly. In this manner, paired hard samples (P^H and P^I) are produced. We then feed the P^H and P^I to the feature encoder ($\mathbb{E}$) and generate the paired features (f^{hard} and f^{inter}).

As an integral aspect of our design, our objective is to produce latent features (f^{latent}) regarding the *hard instance* which are quite perplexing and confusing concerning the *interference class*. Thus, besides providing the hard features (f^{hard}) of the hard instance (P^H) to the generator ($\mathbb{G}$), we treat the features (f^{inter}) from the interference instance (P^I) as the supplemental information and feed them to $\mathbb{G}$, as demonstrated in Fig. 2. Then the generated latent features f^{latent} from $\mathbb{G}$ are therefore extremely confusing and difficult to discriminate with regard to P^H and P^I.

Additionally, to aid the learning process of feature generator $\mathbb{G}$, an "ambiguous label" is further added to the *hard instance* for the purpose of ensuring that f^{latent} are ambiguous and sufficiently challenging. The following is the explanation for "ambiguous label". Typically, a one-hot vector is utilized for the representation of the ground-truth label. To be specific, in the one-hot vector of the category j, the j_{th} element is assigned 1 and the remaining places are all assigned 0. In contrast to the use of one-hot label, we represent the "ambiguous label" as a vector y^{amb}, in which two elements corresponding to the ground-truth category of the *hard instance* and its *interference class* are assigned a value of 0.5 each, with all other elements being set to 0.

The use of "ambiguous label" (y^{amb}) can be treated subsequently as a limitation that makes the generated latent features (f^{latent}) ambiguous in regard to these two classes. Following is a formulation of this constraint:

$$\mathcal{L}_{amb}^{\mathbb{G}} = -\sum_{k=1}^{K} y_k^{amb} \cdot \log \hat{y}_k^{latent} \tag{1}$$

where K represents an aggregate number of active categories, and $\hat{y}^{latent}$ is generated by the class discriminator which processes the generated latent features (f^{latent}) to perform classification.

Eq. (1) assures all generated potential features are sufficiently obscure. Nonetheless, as previously stated, f^{latent} needs to still be credible with inherent information preserved simultaneously. For this purpose, a real-or-fake restraint is applied on f^{latent} to ensure its plausibility, along with a mean-absolute-error restraint to force f^{latent} close to the f^{hard}.

Eq. (2) is the formulation of the mean-absolute-error restraint ($\mathcal{L}_{con}^{\mathbb{G}}$) with the aim of reducing the gap between f^{latent} and f^{hard}. Eq. (3) is the formulation of the real-or-fake restraint ($\mathcal{L}_{rof}^{\mathbb{G}}$), introduced by the RealOrFake Discriminator ($\mathbb{D}^{rof}$) as a measure to ensure that two types of features (the original features (f^{hard}) and the generated features (f^{latent})) reside within the same feature domain.

$$\mathcal{L}_{con}^{\mathbb{G}} = ||f^{latent} - f^{hard}||_1 \tag{2}$$

$$\mathcal{L}_{rof}^{\mathbb{G}} = E[\log \mathbb{D}^{rof}(f^{hard})] + E[\log[1 - \mathbb{D}^{rof}(f^{latent})]] \tag{3}$$

Finally, combining Eq. 2, Eq. 3 and Eq. 1, we can formulate the objective function for our generator ($\mathbb{G}$) as follows:

$$\mathcal{L}^{\mathbb{G}} = \mathcal{L}_{con}^{\mathbb{G}} + \lambda_1 \mathcal{L}_{rof}^{\mathbb{G}} + \lambda_2 \mathcal{L}_{amb}^{\mathbb{G}} \tag{4}$$

Class Discriminator. In order to achieve high discrimination power, we propose a class discriminator ($\mathbb{D}^{cls}$) that can differentiate the latent features (f^{latent}) produced by each *hard instance* from the *interference class*. As part of our class discrimination learning process, we apply a classification constraint ($\mathcal{L}^{\mathbb{D}^{cls}}$) on $\mathbb{D}^{cls}$ to encourage it to assign the proper label (y) of the initial *hard instance* on the basis of the baffling latent features (f^{latent}):

$$\mathcal{L}^{\mathbb{D}^{cls}} = -\sum_{k=1}^{K} y_k \cdot \log \hat{y}_k^{latent} \tag{5}$$

Subsequently, as adversarial learning proceeds, the generated latent features (f^{latent}) that represent the original *hard instance* get increasingly confusing in terms of its *interference class* (i.e., consisting of fewer and fewer differentiation details for $\mathbb{D}^{cls}$ to distinguish classes). Nonetheless, the more ambiguous latent features (f^{latent}) further augment the capability of $\mathbb{D}^{cls}$ to understand the remaining minor discriminative information in the generated latent features f^{latent} in order to differentiate it from the corresponding *interference class*, i.e., $\mathbb{D}^{cls}$ gains increasing power in extracting the relatively subtle discrimination information that is required for better class classification.

Notably, in addition to integrating f^{latent} to train $\mathbb{D}^{cls}$, we also feed the original features (f^{ori}) of original samples into $\mathbb{D}^{cls}$ during adversarial learning, which is illustrated in Fig. 2. It follows that the objective function below would also be applicable to the learning of $\mathbb{D}^{cls}$

$$\mathcal{L}_{ori}^{\mathbb{D}^{cls}} = -\sum_{k=1}^{K} y_k \cdot \log \hat{y}_k^{ori} \tag{6}$$

Previously, we described the adversarial learning scheme as retaining the original features and generating new ones within the same domain. This kind of training scheme that combines Eq. (5) and (6 allows for stabilizing the training of the overall network, thereby providing an effective $\mathbb{D}^{cls}$ to extract minor discrimination information from both the f^{latent} as well as the f^{ori} to distinguish classes. As a result, the derived class discriminator $\mathbb{D}^{cls}$, which has a great deal of power to extract subtle discrimination information and thus efficiently classify the *hard instances* from its corresponding *interference classes*, is able to function as the ultimate activity prediction classifier.

Algorithm 1: HARDer-Net

Input: Partial activity sequences (P) and ground-truth labels (c^{τ})

while *not converge* **do**
- **Backbone learning and HG Bank Filling**
  - Calculate f^{ori} by $\mathbb{E}$;
  - Calculate $\hat{y}^{ori}$ by $\mathbb{D}^{cls}$;
  - Calculate $\mathcal{L}_{ori}^{\mathbb{D}^{cls}}$ with Eq. (6);
  - Update $\mathbb{E}$ and $\mathbb{D}^{cls}$;
  - **if** *rank-1*($\hat{y}^{ori}$) $! = c^{\tau}$ **then**
    - $P^H \leftarrow P$;
    - $c^I \leftarrow$ *rank-1*($\hat{y}$);
    - $\mathbb{H} \leftarrow \{P^H; c^I\}$;
  - **end**
- **end**
- **Adversarial HARDer-Net Learning**
  - Freeze $\mathbb{E}$;
  - Adaptively select and sample P^H and P^I by $\mathbb{H}$;
  - Calculate f^{hard} and f^{inter} by $\mathbb{E}$;
  - Calculate f^{latent} by $\mathbb{G}$;
  - Calculate $\mathcal{L}^{\mathbb{D}^{cls}}$ and $\mathcal{L}^{\mathbb{D}^{rof}}$;
  - Freeze $\mathbb{G}$; Update $\mathbb{D}^{rof}$ and $\mathbb{D}^{cls}$;
  - Calculate $\mathcal{L}^{\mathbb{G}}$;
  - Freeze $\mathbb{D}^{rof}$ and $\mathbb{D}^{cls}$; Update $\mathbb{G}$;
  - Freeze $\mathbb{G}$ and $\mathbb{D}^{rof}$; Update $\mathbb{D}^{cls}$ and $\mathbb{H}$;
- **end**

end

3) Hardness-Guided Bank (HG Bank): In our framework, rather than randomly choosing a hard pair for feature encoding, we propose a Hardness-Guided bank ($\mathbb{H}$) to capture the benefit information obtained from each selection of *hard pairs* in the process of training. Specifically, we utilize a Reinforcement Learning framework to adjust each pick of *hard pair* by calculating the corresponding cumulative reward, as illustrated in the top half of Fig. 2. Consequently, our model is able to identify the exact categories into which hard partial activity sequences may be simply mispredicted.

Fig. 2 illustrates the elementary network structure, which is composed of a feature encoder $\mathbb{E}$ that extracts features from the partially-observed activity sequence of the experiment, and a classifier (referred to as class discriminator in Fig. 2) that is responsible for the task of prediction. Specifically, the encoder firstly processes each partial activity sequence (P) to extract its original features f^{ori}. The original features f^{ori} are then used by the class discriminator to determine the prediction scores $\hat{y}$. If the class discriminator incorrectly predicts the class of the partial sequence instance (P) with prediction scores $\hat{y}$, we treat the activity class c_{r_1} obtaining the rank-one prediction score in $\hat{y}$ as the desired *interference class* (c^I) with regard to P, since c_{r_1} contains the most ambiguous details regarding P. The incorrectly predicted partial activity sequence (P) with inadequate discrimination information is defined as the *hard instance* (P^H) that can be assembled with c^I into a *hard pair*. Obtained *hard pairs* will then be deposited into the HG bank, which is shown in Fig. 2.

With plenty of *hard pairs* collected, HG bank further presents a DQN algorithm [15], [16] to adaptively select *hard pairs* for adversarial learning. Specifically, in this design, our *state* s is defined as the mean value of hidden states at the last layer of the RealOrFake Discriminator ($\mathbb{D}^{rof}$). This is because that in the DQN algorithm, the *state* s should aid in selecting actions that maximize final rewards. The RealOrFake Discriminator is designed for encouraging the feature generator to produce ambiguous yet hard-to-distinguish latent features. And the latent features can encourage the Class Discriminator to achieve higher prediction performances. Thus, the features of the RealOrFake Discriminator, that are directly related to the model accuracy, can be used as the *state* s. Next, since we consider the selection of hard pairs as a decision-making process, we then represent the choosing of *hard pairs* as our *action* a, and following previous methods [15], [16] we utilize the ϵ-greedy policy to balance the exploration and exploitation. Moreover, to boost the performance of our class discriminator, we focus on the prediction accuracy of the Class Discriminator ($\mathbb{D}^{cls}$) and apply it to illustrate the extent to which the action has improved the discrimination performance. Therefore, the prediction accuracy, which aims to be directly boosted, is set as our *reward* r for each training iteration.

Therefore, conditioned on the current *state* s of the RealOrFake Discriminator, our HG bank is able to estimate the $\mathcal{Q}$ value. Then following typical DQN [15], [16], we use ϵ-greedy policy to generate *action* a, i.e., to select the most informative hard pairs from the HG bank to maximize the final reward which is the prediction accuracy. The process is formulated as Eq. 7. And through this conjecture of the future, our HARDer-Net can learn the action-value function $\mathcal{Q}^*$ which corresponds to the optimal policy, and the mean absolute error constraint is formulated in Eq. 8:

$$\mathcal{Q}^{\pi}(s,a) = \mathbb{E}_{s'}[\mathcal{R}(s,a) + \gamma \max_{a'}(\mathcal{Q}^{\pi}(s',a'))|(s,a)] \tag{7}$$

$$\mathcal{L}_{con}^{\mathbb{H}} = ||\mathcal{Q}^*(s,a|\theta) - (\mathcal{R} + \gamma \max_{a'}(\overline{\mathcal{Q}}^*(s',a'))|(s,a))||_1 \tag{8}$$

where $\overline{\mathcal{Q}}$ represents the target $\mathcal{Q}$ function, the parameters of which are intermittently updated based on the latest θ, thus stabilizing the learning process.

C. Implementation Details

In the HARDer-Net training cycle, two phases are involved, specifically backbone training with HG bank augmentation and adversarial learning.

Backbone training & HG bank filling. Fig. 2 illustrates our network's main elements: a class discriminator $\mathbb{D}^{cls}$ and an encoder $\mathbb{E}$. Eq. (6) can be used as a basis for training this backbone. Initially, an encoder $\mathbb{E}$ extracts a mini-batch of original partial activity sequences, denoted as P, with a batch size of B to fill the HG bank. Afterward, the class discriminator calculates predicted scores that function as a

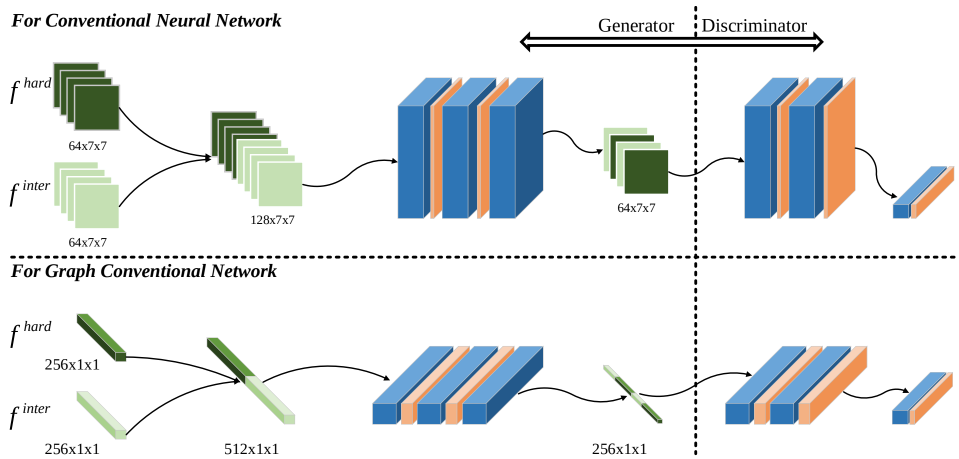


Fig. 3. Detailed structure of Generator and Discriminator. The first row demonstrates the detailed structure of the CNN-based backbone, and the second row demonstrates the detailed structure of the GCN-based backbone.

metric for depositing *hard pairs* into our HG bank, for example, if a sample is incorrectly predicted, the sample with its incorrect prediction class is collected together as a *hard pair* that will be subsequently stored in our HG bank.

Adversarial learning scheme. Prior to adversarial learning, we first freeze the encoder $\mathbb{E}$. Then we sample rB *hard pairs* based on learned policy from the HG bank, in which $0 < r \leq 1$. If no sufficient pairs in the HG bank, we sample and repeat all pairs in the HG bank to reach rB. Conversely, if there are enough pairs, we apply the first-in-first-out scheme to choose rB *hard pairs* from the HG bank. Conditioned on the *interference class* associated with each chosen *hard pair*, we randomly sample an instance belonging to this *interference class* as P^I. After that, P^H and P^I are processed through the encoder $\mathbb{E}$ for feature extraction. The encoded features f^{hard} and f^{inter} are subsequently input into the generator $\mathbb{G}$ to obtain latent features f^{latent}. After obtaining the latent features f^{latent}, we send them into two discriminators, i.e., the RealOrFake discriminator $\mathbb{D}^{rof}$ and the class discriminator $\mathbb{D}^{cls}$, together with ground-truth action labels for the purpose of updating the model parameters. At last, the f^{latent} and the ambiguous label are utilized for updating $\mathbb{G}$. The overall training process is demonstrated in Alg. 1.

Testing. The red arrows presented in Fig. 2 indicate that a segmented skeleton sequence is fed to our encoder for feature extraction during the inference phase. We afterward input these features to $\mathbb{D}^{cls}$ to predict the activity. Considering that $\mathbb{D}^{cls}$ is capable of exploiting minute discrimination information that plays a significant role in differentiating hard samples from corresponding interference classes, the proposed network is capable of achieving an unprecedented level of accuracy in 3D early activity prediction tasks.

IV. Experiments

Evaluations of the proposed method are conducted on the NTU RGB+D dataset [4], the First Person Hand Action (FPHA) dataset [58], the SYSU 3D Human-Object Interaction (HOI) dataset [59] and the UCF-101 dataset [60]. Detailed experiments are conducted on these four datasets as follows.

- **NTU RGB+D** dataset contains a vast collection of data on 3D action recognition and prediction, which has been applied to many applications. From 60 categories of activities, the dataset comprises over 56 thousand videos and more than 4 million frames. It features human skeletons, each consisting of 25 joints depicted in a three-dimensional format. As a result of the large number of baffling samples at the beginning of the activity sequences, this dataset presents a substantial challenge for 3D early action prediction. Two standard evaluation protocols are available in the NTU RGB+D dataset: the Cross Subject evaluation protocol (CS) and the Cross View evaluation protocol (CV). In our experiment, we apply the Cross Subject protocol following existing work [13] by assigning 20 subjects for training and the remaining 20 for testing.
- **First Person Hand Action (FPHA)** dataset [58] represents a challenging dataset of 3D hand gestures. There are six subjects represented in the dataset, each capturing first-person hand activities involving interactions with 3D objects. It comprises an extensive collection of over 100,000 frames, spanning 45 unique categories of hand activities. An individual hand skeleton consists of 21 joints and is characterized using 3D coordinates. We assess the effectiveness of our framework using the FPHA dataset conforming to the standard evaluation protocol as [58], involving 600 600 training and 575 testing activity sequences.
- **SYSU 3D Human-Object Interaction (HOI)** dataset is a widely recognized RGB-D activity dataset that focuses on human-object interactions. In the dataset, twelve different activities are assigned to 40 subjects, and participants operate one of six different objects for each activity: besom, phone, wallet, chair, bag, and mop. On SYSU 3DHOI we investigate our method according to [61] and sequences executed by one-half of the subjects are utilized for learning the model parameters, while sequences executed by the remaining half serve to test the model.
- **UCF-101** dataset is a challenging and unconstrained

TABLE I

QUANTITATIVE RESULTS (%) COMPARISON ON THE NTU RGB+D DATASET (CROSS-SUBJECT). OUR METHOD OUTPERFORMS THE BACKBONE MODEL ("W/O HARDER-NET") SIGNIFICANTLY. FURTHERMORE, IT OUTPERFORMS STATE-OF-THE-ART 3D EARLY ACTIVITY PREDICTION METHODS BY A WIDE MARGIN. REFER TO FIG. 4 FOR VISUALIZATION.

Methods	Observation Ratios					
	20%	40%	60%	80%	100%	AUC
Ke *et al.* [23]	8.34	26.97	56.78	75.13	80.43	45.63
Jain *et al.* [20]	7.07	18.98	44.55	63.84	71.09	37.38
Aliakbarian *et al.* [21]	27.41	59.26	72.43	78.10	79.09	59.98
Wang *et al.* [2]	35.85	58.45	73.86	80.06	82.01	60.97
Pang *et al.* [66]	33.30	56.94	74.50	80.51	81.54	61.07
Weng *et al.* [3]	35.56	54.63	67.08	72.91	75.53	57.51
Ke *et al.* [13]	32.12	63.82	77.02	82.45	83.19	64.22
Li *et al.* [67]	38.18	71.19	82.25	86.33	87.20	-
Wang *et al.* [68]	42.53	**72.64**	83.12	86.75	87.21	70.67
w/o HARDer-Net	37.82	67.87	79.22	83.39	84.52	66.91
HARDer-Net w/o RL	42.39	72.24	82.99	86.75	87.54	70.56
HARDer-Net	**43.22**	72.43	**83.17**	**87.00**	**87.80**	**70.87**

RGB video-based dataset widely used for the understanding of human action, pose, and behavior. The dataset comprises a total of 13,320 full videos, encompassing 101 distinct action classes that have been categorized into 101 content-based categories. The entire video clip collection consists of over 27 hours, categorized into five distinct types (body movement, human-human interaction, human-object interaction, playing instruments, and sports). For UCF-101, we employ the same setting as [62] by training with the initial 15 groups of videos, conducting validation using the subsequent three groups, and finally, performing testing on the remaining videos.

Evaluated Models. To assess the effectiveness of our method, we consider three variants, specifically "w/o HARDer-Net", "HARDer-Net w/o RL" and "HARDer-Net". (1) "w/o HARDer-Net": In fact, here is the backbone model of our network, which is composed of the feature encoder and the classifier; (2) "HARDer-Net w/o RL": Here is the proposed 3D early activity prediction model (HARDer-Net) with the Hardness-Guided bank. However, the *hard pairs* are randomly selected from the HG Bank, i.e., HARD-Net in our previous conference paper; (3) "HARDer-Net": Here is our proposed 3D early activity prediction model with an elaborate Hardness-Guided bank structure that further enhances the performance of our framework for 3D early action prediction. Note that here the *hard pairs* are selected by using our reinforcement learning scheme.

To evaluate the HARDer-Net, we build our proposed method over two baseline encoders, specifically, CNN [63] and GCN [27], corresponding to Tab. V. Detailed descriptions of both baseline encoders are provided in their respective papers [27], [63]. In addition, we follow Radford *et al.* [64] in designing our generator and RealOrFake discriminator, and the class discriminator is implemented on the strength of multi-layer perceptron. Moreover, The weights λ_1 and λ_2 in Eq. (4) are both set to 1.

The experiments are all performed using the Pytorch framework with a single GeForce RTX 3080 Ti GPU. We set the batch size B to be 128. Adam [65] optimizer is utilized in the training of our end-to-end network with the initial learning rate set to 2×10^{-4}. For the highly large NTU RGB+D dataset and UCF-101 dataset, we set the Hardness-Guided bank size to 5000, and for the small FPHA dataset and SYSU 3DHOI dataset, we set it to 100. Every time the network learning algorithm is run, an appropriate ratio of r $(4:1)$ is established with the original instances and the *hard pair* instances utilized in the learning of our network.

Network Architecture. To generate latent features (f^{latent}) from *hard features* (f^{hard}) and *interference features* (f^{inter}), a feature generator is designated to investigate the relationship between *hard instances* and corresponding *interference classes*. The remarkable thing is GCN and CNN backbones generate widely different features. Consequently, two similarly constructed deep networks are proposed in Fig. 3, where blue blocks represent convolutional layers with 1×1 kernel size (for CNN backbone) and fully connected layers (for GCN backbone), respectively. Meanwhile, orange blocks represent the non-linear activation functions. Take the experiment on the NTU-RGB+D dataset as an example: in our generator, we feed the concatenated $f^{hard} \in \mathbb{R}^{64\times7\times7}$ and $f^{inter} \in \mathbb{R}^{64\times7\times7}$ into convolutional layers to obtain the corresponding $f^{latent} \in \mathbb{R}^{64\times7\times7}$. In the next step, we incorporate f^{latent} as input to our discriminator to enhance its capability to distinguish them from their corresponding *interference class*. HG bank is identified to map the last layer of hidden states $f^{state} \in \mathbb{R}^{256}$ in our RealOrFake discriminator into the action $f^{action} \in \mathbb{R}^{5120}$ for selection of preserved hard pairs. Note that in both architectures, f^{hard} and f^{inter} are concatenated and then processed by the feature generator in order to achieve the latent features f^{latent} that have the same shape as f^{hard} and f^{inter}.

In our HARDer-Net, the proposed HG bank aims to intelligently sample the stored *hard pairs* in the original bank space. As the adversarial learning phase progresses, HG bank selects feature pairs that provide more valid information for our discriminator to lift its discrimination capacity. Namely, HG bank successfully reduces the interference of futile information to model training caused by random selection.

A. Experiments on the NTU RGB+D Dataset

First, we make a comparison of the proposed HARDer-Net against the state-of-the-art approaches utilizing the NTU RGB+D dataset. Results of the Cross Subject protocol involving different observation ratios are presented in Tab. I and Fig. 4. As illustrated in Tab. I, our proposed HARDer-Net consistently shows the highest performance across all observation ratios, demonstrating the efficiency of HARDer-Net. Especially when the observation ratio is low, our method outperforms the state-of-the-art work and the backbone model significantly. As the significant improvements indicate, our approach is effective for detecting subtle but meaningful distinctions concerning discrimination.

Furthermore, we also apply the area under the curve, abbreviated AUC, to estimate the comprehensive performance

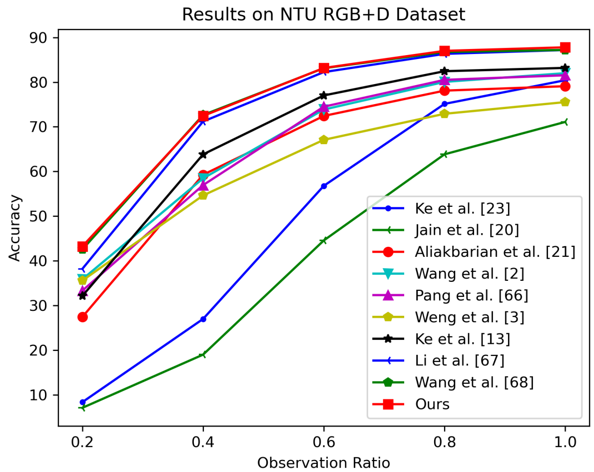


Fig. 4. An analysis of the performance of 3D early activity prediction task on NTU RGB+D datasets. A large margin of improvement is achieved by our method over existing methods.

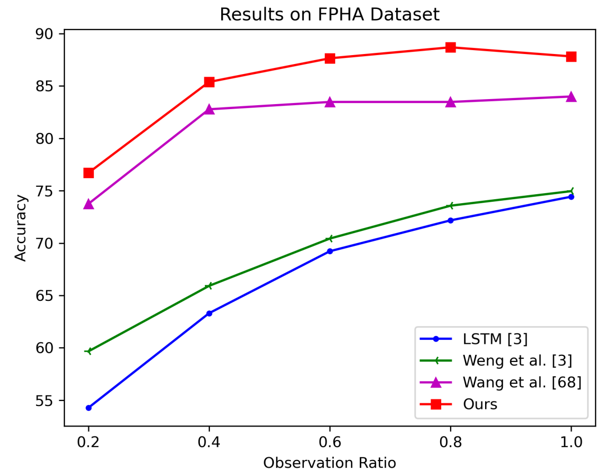


Fig. 5. An analysis of the performance of 3D early activity prediction task on FPHA datasets.

TABLE II
QUANTITATIVE RESULTS (%) COMPARISON ON THE FPHA DATASET WITH STATE-OF-THE-ARTS. REFER TO FIG. 5 FOR VISUALIZATION.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
LSTM [3]	54.26	63.30	69.22	72.17	74.43	64.11
Weng *et al.* [3]	59.65	65.91	70.43	73.57	74.96	66.66
Wang *et al.* [68]	73.74	82.78	83.48	83.48	84.00	77.12
w/o HARDer-Net	62.26	74.61	79.65	82.09	83.48	72.17
HARDer-Net w/o RL	71.83	82.78	86.09	87.13	87.30	78.56
HARDer-Net	**76.70**	**85.39**	**87.65**	**88.70**	**87.83**	**80.71**

of our proposed HARDer-Net, measuring the average precision across various observation ratios following [3], [21], [66]. Additionally, as evidenced in Tab. I, our method attains the highest average AUC score of 70.87%, when compared with existing methods as well as the backbone model ("w/o HARDer-Net"). It is worth noting that our HARDer-Net exceeds the backbone model by a margin of 3.96%, demonstrating that the constructed adversarial learning scheme is effective at understanding and perceiving subtle differences inside of *hard classes* and assisting the class discriminator in discriminating *hard instances*. It's noteworthy that when compared with "HARDer-Net w/o RL", our full HARDer-Net still outperforms by a notable margin. This demonstrates that using the proposed RL-based selecting scheme, the selection of *hard pairs* is driven by the reward, i.e., the recognition accuracy. And this further explicitly boosts the recognition performances of our HARDer-Net.

B. Experiments on the FPHA Dataset

As part of our investigation of the competency of our proposed HARDer-Net on 3D gesture datasets, we conduct comprehensive experiments on the FPHA dataset. As shown in Fig. 5 and Tab. II, our proposed HARDer-Net outperforms Weng *et al.* [3] consistently in all ranges of observation ratios.

TABLE III
QUANTITATIVE RESULTS (%) COMPARISON ON THE SYSU 3DHOI DATASET WITH STATE-OF-THE-ARTS. REFER TO FIG. 6 FOR VISUALIZATION.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
Jain *et al.* [69]	31.61	53.37	68.71	73.96	75.53	57.23
Ke *et al.* [70]	26.76	52.86	72.32	79.40	80.71	58.89
Kong *et al.* [71]	51.75	58.83	67.17	73.83	74.67	61.33
Ma *et al.* [72]	57.08	71.25	75.42	77.50	76.67	67.85
Aliakbarian *et al.* [73]	56.11	71.01	78.39	80.31	78.50	69.12
Hu *et al.* [14]	56.67	75.42	80.42	82.50	79.58	71.25
Ke *et al.* [74]	58.81	74.21	82.18	84.42	83.14	72.55
Wang *et al.* [62]	63.33	75.00	81.67	86.25	87.92	74.31
Li *et al.* [67]	63.46	80.93	**87.92**	**90.38**	**90.47**	-
Wang [68]	65.00	81.67	86.67	89.17	89.25	**78.01**
w/o HARDer-Net	62.92	80.83	85.42	87.08	87.50	76.50
HARDer-Net w/o RL	63.75	81.25	85.83	87.92	87.92	77.06
HARDer-Net	**65.00**	**81.67**	86.25	88.33	88.33	77.59

As compared to the baseline model, with very low observation ratios and inadequate discrimination information in the early stages, our HARDer-Net achieves the most notable performance gains by 14.44% at the 20% observation ratio and 10.78% at the 40% observation ratio, since it is capable of mining minor discrepancies.

Another observation is that the AUC score of action prediction decreases at the ending stages. One possible explanation for this issue is that some frames at the end of the skeleton sequence contain postures and motions that have little relation to the class label of the current action.

C. Experiments on the SYSU 3DHOI Dataset

A comprehensive study has been conducted using a trendy RGB-D activity dataset named SYSU 3DHOI to illustrate the effectiveness of the HARDer-Net respecting the 3D early action prediction problem. As presented in Tab. III and Fig. 6, our proposed HARDer-Net yields the highest performance

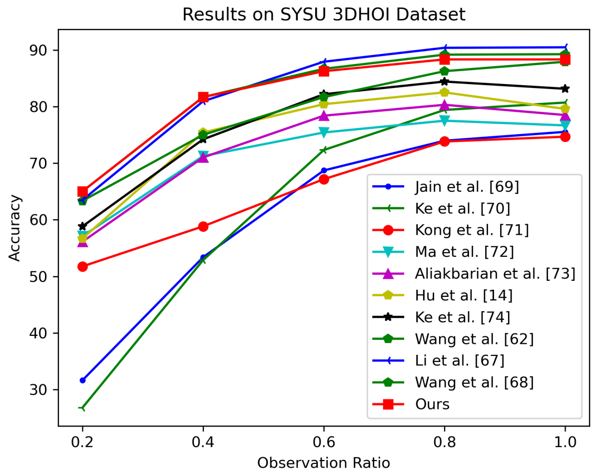


Fig. 6. An analysis of the performance of 3D early activity prediction task on SYSU 3D HOI datasets.

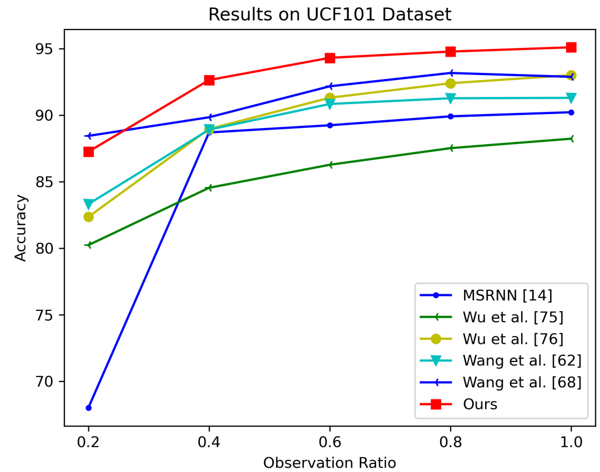


Fig. 7. An analysis of the performance of 3D early activity prediction task on UCF-101 datasets.

TABLE IV
QUANTITATIVE RESULTS (%) COMPARISON ON THE UCF-101 DATASET. REFER TO FIG. 7 FOR VISUALIZATION.

	Observation Ratios					
Methods	10%	30%	50%	70%	90%	AUC
MSRNN [14]	68.01	88.71	89.25	89.92	90.23	80.89
Wu *et al.* [75]	80.24	84.55	86.28	87.53	88.24	80.57
Wu *et al.* [76]	82.36	88.97	91.32	92.41	93.02	84.66
Wang *et al.* [62]	83.32	88.92	90.85	91.28	91.31	84.27
Wang *et al.* [68]	**88.45**	89.85	92.18	93.18	92.89	86.22
w/o HARDer-Net	83.19	91.12	93.29	93.97	94.21	86.20
HARDer-Net w/o RL	84.19	91.46	93.47	94.32	94.77	86.62
HARDer-Net	87.26	**92.65**	**94.32**	**94.79**	**95.11**	**87.72**

among all of the observation ratios compared with Wang *et al.* [62].

Furthermore, due to the ability of our HARDer-Net to extract minor discrepancies in contrast with the baseline model under low observation ratios and lacks sufficient discrimination information, it further enhances the early prediction performance with the gain of 2.08% at 20% observation ratio. The performance gains show that driven directly by the reward, the selected hard pairs enable our recognition model to exploit the minor yet significant differences, which can further improve the recognition accuracy.

We also note that an increase in the observation ratio from 80% to 100% does not correspond with an improvement in the prediction accuracy of the proposed HARDer-Net model. A potential reason is that the model is already overfitted when the observation ratio is at 80% owing to the limitation of the dataset size, leading to more data observations that no longer raise the performance of our HARDer-Net.

D. Experiments on the UCF-101 Dataset

To extensively evaluate the proposed HARDer-Net on the 3D early action recognition dataset, we evaluate our model on five different observation ratios (10%, 30%, 50%, 70%, 90%) on the UCF-101 dataset. Comparisons with existing approaches [14], [62], [75], [76] are presented Tab. IV for various observation ratios.

The comparison results show that by using the HG bank to adaptively select hard pairs for adversarial learning, a significant improvement has been achieved in the 3D early activity prediction performance of our HARDer-Net. In this regard, the novel method we propose has proven to be highly effective in lessening the interference of insignificant information to model training caused by random selection of preserved hard pairs.

Additionally, as shown in Tab. IV, the accuracy of 100% observation ratio outperforms the accuracy of 80% observation ratio, which indicates that for the UCF101 dataset, when number of observed frames increases, more discriminative human motion details are revealed, thereby enhancing the prediction performances of 100% observation ratio. However, when compared to SYSU 3DHOI dataset (shown in Tab. III), we observed that performances for 80% observation ratio and 100% observation ratio are identical. We presumed that the final 20% video frames in the SYSU 3DHOI dataset possibly do not contribute sufficient representative discrimination information in predicting human actions.

It is also worth noting that, for both UCF101 and SYSU 3DHOI datasets, the proposed HADRer-Net achieves promising prediction accuracy across all observation ratios when compared with existing approaches. This demonstrates that our reward-driven HG bank mechanism is able to adaptively capture the representative subtle cues for different datasets containing heterogeneous characteristics.

E. Ablation Study

This section presents an extensive ablation study based on the NTU-RGB+D dataset to verify the best setting for our proposed model's components, following existing works [2], [3], [13], [66] in the early activity prediction community.

Impact of Bank Size. We proceed to assess the performance of the HG bank across a range of bank sizes. In HARDer-Net,

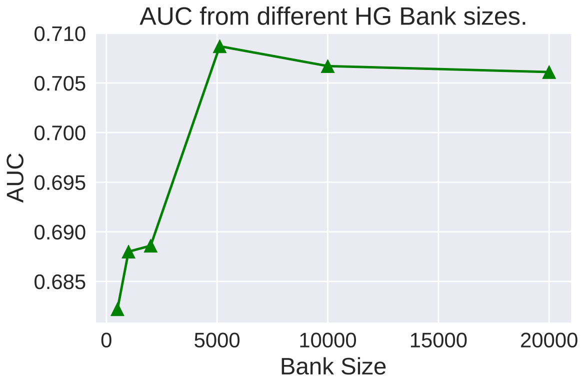


Fig. 8. A comparison of the impacts of different HG bank sizes. Generally, when the size of our proposed HG bank increases from 1000 to 5000, our model's performance increases and peaks, but further expansion of the bank size does not enhance its performance.

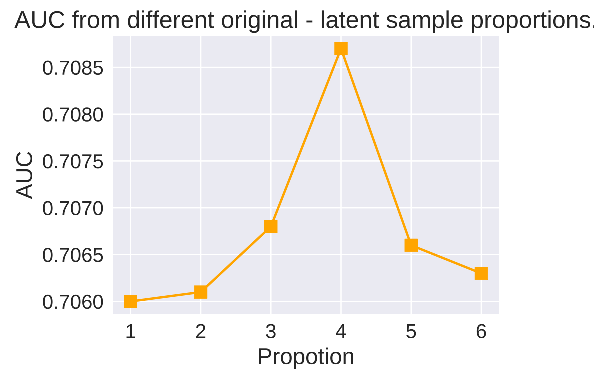


Fig. 9. A comparison of the impacts of varying the proportion of original samples and *hard pair* samples in HG bank used to train networks. It is shown that our model achieves the highest AUC score when the proportion of the original sample size to the *hard pair* sample size is 4:1.

TABLE V
PERFORMANCE GAIN (%) ON NTU RGB+D DATASET (CROSS-SUBJECT) BROUGHT BY OUR HARDER-NET WITH DIFFERENT BACKBONES.

		Observation Ratios				
Backbone	**Methods**	20%	40%	60%	80%	100%
CNN backbone [13]	w/o HARDer-Net	34.01	63.16	75.87	81.39	82.24
	HARDer-Net	**36.52**	**65.63**	**77.81**	**82.88**	**83.98**
	Δ	**+2.51**	**+2.47**	**+1.94**	**+1.49**	**+1.74**
GCN backbone [27]	w/o HARDer-Net	37.82	67.87	79.22	83.39	84.52
	HARDer-Net	**43.22**	**72.43**	**83.17**	**87.00**	**87.80**
	Δ	**+5.40**	**+4.56**	**+3.95**	**+3.61**	**+3.28**

HG bank takes responsibility for storing hard pairs generated in the training process and sampling suitable items for the adversarial learning scheme. If the size of our proposed HG bank is not large enough, it will not hold sufficient information for adversarial learning; conversely, if we set the bank with a large size, it will contain a quantity of irrelevant items, thus reducing the efficiency of HG bank's sampling. A representation of the results can be found in Fig. 8. The AUC, a metric quantifying average precision across all observation ratios, exhibits a rapid increase from a small to a large bank size, before stabilizing at a sufficiently large size (e.g., size 5000). In this case, the additional performance gain is restricted when the inherent threshold is reached, due to the total number of *hard pairs* in the dataset.

Impact of proportions between original features and latent features for training. We find that the optimum ratio for original features and the latent features is 4:1 in our experimental results displayed in Fig. 9.

The reason for this is that if we utilize excessive amounts of original features for network training, our adversarial learning scheme will extract less meaningful information for better discrimination. The use of too many latent features for training may however reduce the performance of our HARDer-Net relative to the original samples. Furthermore, small differences in the performance of various ratios (1:1 to 6:1) suggest that the HARDer-Net does not exhibit sensitivity to ratios. As shown in Fig. 9, our HARDer-Net achieves AUCs that are within a narrow range (70.6% to 70.9%), demonstrating its robustness against ratios. Furthermore, it is noteworthy that all of these AUC values surpass the baseline performance of 66.9% by a substantial margin, which serves as strong proof of the effectiveness of our HARDer-Net.

Impact of Backbone Encoder. Tests of our framework have been extensively conducted on CNN and GCN backbones and the proposed approach has been shown to be effective. As shown in Tab. V, In both backbone models, our HARDer-Net enhances early prediction performance, especially when observation ratios are extremely low. The results of this study indicate that our HARDer-Net has the capacity to exploit relatively minor discrimination information for the purpose of 3D early activity prediction.

However, we would like to clarify that our previous conference submission, HARD-Net, is established on 2S-AGCN backbone [27], while the most recent works conduct their experiments on MS-G3D backbone [77] which is a more powerful GCN. Therefore, to make a fair comparison, we replace the 2S-AGCN [27] in our HARDer-Net with MS-G3D [77] and the performances are shown in Tab. VI. The experimental results demonstrate that our HARDer-Net achieves state-of-the-art performances when compared with the most recent works using the same backbone network, which further demonstrates the efficacy of our HARDer-Net.

TABLE VI
QUANTITATIVE RESULTS (%) COMPARISON ON THE NTU RGB+D DATASET (CROSS-SUBJECT) USING **MS-G3D** AS BACKBONE FEATURE ENCODER.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
ERA [78]	53.98	74.34	85.03	88.35	88.45	73.87
UPS [79]	53.25	75.06	85.35	-	-	-
Magi-Net [80]	46.68	75.11	84.87	88.12	88.72	72.77
TODO-Net [81]	45.95	74.37	84.61	87.71	88.62	72.32
HARDer-Net	54.11	75.03	85.40	88.71	88.74	74.24

Impact of choice of states for the HG bank. To search for the optimal states for the reward-driven HG bank, we design three different states: (1) s_1 is defined as the mean value of hidden states at the last layer of the Class Discriminator $\mathbb{D}^{cls}$; (2) s_2 is defined as the mean value of the latent features; (3) s_2 is defined as the mean value of hidden states at the last layer of the RealOrFake Discriminator $\mathbb{D}^{rof}$. The results are shown in Tab. VII. As we can see in Tab. VII, compared with our previous conference submission, i.e., "HARDer-Net w/o RL", our newly-designed HG bank consistently achieves better performances. Also, when defining the hidden features from the RealOrFake Discriminator $\mathbb{D}^{rof}$ as state s, our HARDer-Net obtains the highest prediction performance. This is possibly because the hidden features from the $\mathbb{D}^{rof}$ contain ambiguous information from the hard pairs which enables our prediction to mine the subtle cues and further improve the accuracy.

TABLE VII
QUANTITATIVE RESULTS (%) COMPARISON ON DIFFERENT STATES FOR HG BANK.

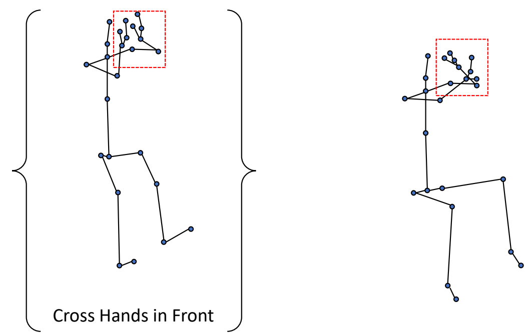
	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
HARDer-Net w/o RL	42.39	72.24	82.99	86.75	87.54	70.56
HARDer-Net w/ s_1	43.01	72.39	83.00	86.80	87.60	70.72
HARDer-Net w/ s_2	42.87	72.40	83.04	86.90	87.61	70.73
HARDer-Net w/ s_2	43.22	72.43	83.17	87.00	87.80	70.87

Impact of ϵ. In typical DQN [15], [16], the ϵ-greedy policy is used to decide whether to select the top-1 action or to randomly explore non-optimal actions, with the purpose of encouraging the robustness of learned models. Therefore, we conduct ablation experiments on how often the model should explore and how often the model should exploit further as shown in the Tab. VIII. In our previous conference submission, i.e., "HARDer-Net", the bank is conducting random exploration every time. When we gradually increase the chance of exploitation, the prediction performances improve accordingly. This means that by explicitly focusing on those informative hard pairs, our model can learn more robust representations that benefit the action prediction. It's also noteworthy that when we linearly decay the chance of exploration (ϵ), our HARDer-Net performs the best. The reason might be that at the beginning stages, the model has not been optimized well thus it needs to explore more samples. As the training process going, the model can easily identify those "easy" samples and it needs to exploit those really hard samples with subtle cues to boost the prediction abilities.

TABLE VIII
QUANTITATIVE RESULTS (%) COMPARISON ON DIFFERENT ϵ SCHEDULING FOR ϵ-GREED POLICY.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
HARDer-Net w/o RL	42.39	72.24	82.99	86.75	87.54	70.56
ϵ = 0.5	42.69	72.14	83.02	86.85	87.64	70.63
ϵ = 0.1	43.20	72.33	83.15	86.98	87.76	70.83
Linear ϵ	43.22	72.43	83.17	87.00	87.80	70.87

Visualization of HG bank selection. As shown in Fig. 10, the sample (a), which is selected by our HG bank, originally belongs to the "Wipe Face" but is wrongly classified into "Cross Hands in Front". The only difference between these two actions lies in the subtle cues of hand gestures. (For better demonstration, the action sample (b) belongs to the "Cross Hands in Front" category.) Therefore, this demonstrates that our reward-driven HG bank focuses on those truly representative hard pairs, which can further encourage the prediction model to exploit the minor yet significant cues to obtain better prediction performances.



(a) Hard Sample and Interference Class (b) Sample from "Cross Hands in Front"

Fig. 10. Qualitative analysis of the hard pairs selected by HG bank.

Efficiency Analysis. The number of parameters increases approximately by 3.2% and the inference time increases by 2% on Nvidia RTX 3080 Ti. This shows that our HARDer-Net achieves much better performances with trivial computational costs increasing.

V. Discussion

In this research, we employ a Hardness-Guided Discrimination Network (HARDer-Net) which iteratively memorizes and exploits *hard pairs* susceptible to inadequate discrimination information. This is achieved through the implementation of an innovative adversarial hardness-guided learning scheme, paired with a Hardness-Guided (HG) Bank. More precisely, the adversarial hardness-guided learning scheme enables the network to discern and extract subtle yet meaningful discrimination information within the feature space, consequently enhancing the precision of predictions. Concurrently, the Hardness-Guided Bank, augmented by a hardness-guided deep reinforcement learning mechanism, refines the selection process of *hard pairs* with a primary focus on optimizing recognition accuracy. As a result, our advanced HARDer-Net exhibits a distinct superiority over existing state-of-the-art models on four challenging datasets, as illustrated in Tables I to IV.

Nonetheless, our proposed HARDer-Net also reveals certain limitations. For instance, within the FPHA dataset, the AUC score for action prediction diminishes in the final stages, potentially due to some frames at the end of the skeleton sequence containing postures and motions unrelated to the class label of the current action. Alternatively, it is plausible that 80% of the skeleton sequence contains sufficient data for our model to render accurate predictions. Moreover, in the SYSU 3DHOI dataset, there is no corresponding growth in prediction accuracy when the observation ratio increases from 80% to 100%, suggesting a potential overfitting issue at higher observation ratios due to the limitations of the dataset.

For future research, it would be beneficial to explore ways to address the identified limitations. One approach could involve refining the sensitivity of our model to the latter stages of activity sequences, thereby ensuring the maintenance of accurate predictions even when the availability of discrimination information diminishes. Furthermore, expanding the datasets or diversifying the data sources could partially mitigate the overfitting issues observed at higher observation ratios.

VI. Conclusion

We have proposed a new Hardness-Guided Discrimination Network (HARDer-Net) for 3D early activity prediction. This network allows explicit probes into the associations between a readily mispredicted instance, called *hard instance*, and its corresponding class into which it is wrongly classified, called *interference class*. Further, an adversarial learning scheme is constructed to extract slight differences within this *hard instance* - *interference class* pair through the generation of ambiguous and less discriminative latent features conditioned upon the given pair to represent original *hard instances*. Besides, a deep reinforcement learning-based HG bank is designed to adaptively select hard pairs from retained pairs for adversarial learning to enhance the performance of our network. Additionally, we construct a class discriminator to differentiate the latent features derived from the corresponding *interference classes*. Taking advantage of such a framework design, HARDer-Net achieves superior performance in comparison with the state-of-the-art approaches on four challenging datasets.

References

[1] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, and J. Lai, "Real-time rgb-d activity prediction by soft regression," in *European Conference on Computer Vision*, 2016, pp. 280–296.

[2] X. Wang, J. Hu, J. Lai, J. Zhang, and W. Zheng, "Progressive teacher-student learning for early action prediction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3551–3560.

[3] J. Weng, X. Jiang, W. Zheng, and J. Yuan, "Early action recognition with category exclusion using policy-based reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.

[4] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2013.

[6] J. Liu, H. Ding, A. Shahroudy, L.-Y. Duan, X. Jiang, G. Wang, and A. C. Kot, "Feature boosting network for 3d pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 494–501, 2020.

[7] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2823–2832.

[8] Y. Cai, L. Ge, J. Cai, N. Magnenat-Thalmann, and J. Yuan, "3d hand pose estimation using synthetic data and weakly labeled rgb images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[9] Z. Huang, Y. Qin, X. Lin, T. Liu, Z. Feng, and Y. Liu, "Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1868–1883, 2023.

[10] X. Sun, H. Sun, B. Li, D. Wei, W. Li, and J. Lu, "Defeenet: Consecutive 3d human motion prediction with deviation feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5527–5536.

[11] Y. Wen, H. Pan, L. Yang, J. Pan, T. Komura, and W. Wang, "Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 21 243–21 253.

[12] Q. Lin, L. Yang, and A. Yao, "Cross-domain 3d hand pose estimation with dual modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 184–17 193.

[13] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Transactions on Image Processing*, vol. 29, pp. 959–970, 2020.

[14] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang, "Early action prediction by soft regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2568–2583, 2018.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[17] T. Li, J. Liu, W. Zhang, and L. Duan, "Hard-net: Hardness-aware discrimination network for 3d early activity prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 420–436.

[18] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 816–833.

[19] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, April 2018.

[20] A. Jain, A. Singh, H. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," *Arxiv*, 2015.

[21] M. Aliakbarian, F. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," 2017.

[22] W. Zhu, C. Lan, J. Xing, Y. Li, L. Shen, W. Zeng, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *AAAI*, 2016.

[23] Q. Ke, M. Bennamoun, S. An, F. A. Sohel, and F. Boussaïd, "A new representation of skeleton sequences for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4570–4579.

[24] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2969–2978.

[25] H. Luo, G. Lin, Y. Yao, Z. Tang, Q. Wu, and X. Hua, "Dense semantics-assisted networks for video action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3073–3084, 2022.

[26] H. Liu, Y. Liu, Y. Chen, C. Yuan, B. Li, and W. Hu, "Transkeleton: Hierarchical spatial–temporal transformer for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4137–4148, 2023.

[27] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[28] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[29] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[31] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Y. Chen, H. Ge, Y. Liu, X. Cai, and L. Sun, "Agpn: Action granularity pyramid network for video action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3912–3923, 2023.

[34] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *European Conference on Computer Vision*, vol. 8693, 2014.

[35] Y. Kong, S. Gao, B. Sun, and Y. Fu, "Action prediction from videos via memorizing hard-to-predict samples," in *AAAI*, 2018.

[36] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3662–3670, 2017.

[37] W. Xu, J. Yu, Z. Miao, L. Wan, and Q. Ji, "Prediction-cgan: Human action prediction with conditional generative adversarial networks," *Proceedings of the ACM International Conference on Multimedia*, 2019.

[38] Y. Kong, Z. Tao, and Y. Fu, "Adversarial action prediction networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 539–553, 2020.

[39] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Predicting the future: A jointly learnt model for action anticipation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5561–5570.

[40] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1453–1467, 2020.

[41] W. Wang, F. Chang, C. Liu, G. Li, and B. Wang, "Ga-net: A guidance aware network for skeleton-based early activity recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1061–1073, 2023.

[42] W. Wang, F. Chang, J. Zhang, R. Yan, C. Liu, B. Wang, and M. Z. Shou, "Magi-net: Meta negative network for early activity prediction," *IEEE Transactions on Image Processing*, vol. 32, pp. 3254–3265, 2023.

[43] W. Guan, X. Song, K. Wang, H. Wen, H. Ni, Y. Wang, and X. Chang, "Egocentric early action prediction via multimodal transformer-based dual action prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4472–4483, 2023.

[44] N. Zheng, X. Song, T. Su, W. Liu, Y. Yan, and L. Nie, "Egocentric early action prediction via adversarial knowledge distillation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1–21, 2023.

[45] I. Loshchilov and F. Hutter, "Online batch selection for faster training of neural networks," 2015.

[46] A. Shrivastava, H. Mulam, and R. Girshick, "Training region-based object detectors with online hard example mining," 2016.

[47] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb 2020.

[48] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440–1448.

[49] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

[50] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[51] P. Felzenszwalb, R. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 1627–45, 09 2010.

[52] M. Cho, T. young Chung, H. Lee, and S. Lee, "N-rpn: Hard example learning for region proposal networks," 2022.

[53] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[54] B. Uzkent, C. Yeh, and S. Ermon, "Efficient object detection in large images using deep reinforcement learning," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1824–1833.

[55] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[56] G. Vecchio, S. Palazzo, D. Giordano, F. Rundo, and C. Spampinato, "Mask-rl: Multiagent video object segmentation framework through reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5103–5115, 2020.

[57] J. Chen, C. Gao, E. Meng, Q. Zhang, and S. Liu, "Reinforced structured state-evolution for vision-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 15 450–15 459.

[58] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[59] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5344–5352.

[60] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[61] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[62] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3556–3565.

[63] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *IJCAI*, 2018, pp. 786–792.

[64] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.06434

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[66] G. Pang, X. Wang, J.-F. Hu, Q. Zhang, and W.-S. Zheng, "Dbdnet: Learning bi-directional dynamics for early action prediction," in *IJCAI*, 2019, pp. 897–903.

[67] G. Li, N. Li, F. Chang, and C. Liu, "Adaptive graph convolutional network with adversarial learning for skeleton-based action prediction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1258–1269, 2021.

[68] W. Wang, F. Chang, C. Liu, G. Li, and B. Wang, "Ga-net: a guidance aware network for skeleton-based early activity recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1061–1073, 2021.

[69] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3118–3125.

[70] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[71] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1473–1481.

[72] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1942–1950.

[73] M. Sadegh Aliakbarian, F. Sadat Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 280–289.

[74] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Transactions on Image Processing*, vol. 29, pp. 959–970, 2019.

[75] X. Wu, R. Wang, J. Hou, H. Lin, and J. Luo, “Spatial–temporal relation reasoning for action prediction in videos,” *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1484–1505, 2021.

[76] X. Wu, J. Zhao, and R. Wang, “Anticipating future relations via graph growing for action prediction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2952–2960.

[77] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[78] L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, “Era: Expert retrieval and assembly for early action prediction,” in *European Conference on Computer Vision*. Springer, 2022, pp. 670–688.

[79] ——, “Unified pose sequence modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 019–13 030.

[80] W. Wang, F. Chang, J. Zhang, R. Yan, C. Liu, B. Wang, and M. Z. Shou, “Magi-net: Meta negative network for early activity prediction,” *IEEE Transactions on Image Processing*, 2023.

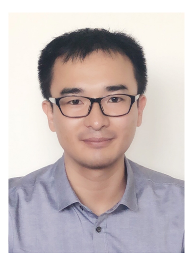
[81] W. Wang, F. Chang, C. Liu, B. Wang, and Z. Liu, “Todo-net: Temporally observed domain contrastive network for 3-d early action prediction,” *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Tianjiao Li is currently a PhD student at SUTD. He got his bachelor’s and master’s degrees from Shandong University in 2016 and 2020 respectively. His research interests include computer vision, action recogntion, early action prediction and pose estimation.

Yang Luo is currently a PhD student at NUS. He completed his master’s degree in Artificial Intelligence from NUS in 2023, after receiving his bachelor’s degree in software engineering from Wuhan University in 2021. His research interests include computer vision and machine learning.

Wei Zhang is currently a Professor with the School of Control Science and Engineering, Shandong University, China. He received the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong in 2010. He has published over 120 papers in international journals and refereed conferences. His research interests include computer vision, image processing, pattern recognition, and robotics. Dr.Zhang served as a program committee member and a reviewer for various international conferences and journals in image processing, computer vision, and robotics.

Lingyu Duan (Member, IEEE) received the Ph.D.degree in information technology from The University of Newcastle, Callaghan, NSW, Australia, in 2008. He is currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, China, and since 2012, he has been the Associate Director of the Rapid-Rich Object Search Laboratory, a joint lab between Nanyang Technological University, Singapore, and Peking University. Since 2019, he has been with Peng Cheng Laboratory, Shenzhen, China. He has authored or coauthored about 200 research papers. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics. He was the Co-Editor of the MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13) and MPEG Compact Descriptor for Video Analytics standard (ISO/IEC 15938-15). He is currently an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, ACM Transactions on Intelligent Systems and Technology and ACM transactions on Multimedia Computing, Communications, and Applications, and the Area Chair of the ACM MM and IEEE ICME. He is a Member of the MSA Technical Committee in IEEE-CAS Society. He was the recipient of the IEEE ICME best paper awards in 2020 and 2019, the IEEE VCIP best paper award in 2019, EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, the China Patent Award for Excellence in 2017, and the National Information Technology Standardization Technical Committee Standardization Work Outstanding Person Award in 2015.



Jun Liu (Senior Member, IEEE) received the B.Eng. degree from Central South University, the M.Sc. degree from Fudan University, and the Ph.D.degree from Nanyang Technological University. His research interests include computer vision and artificial intelligence. He is currently the regular Area Chair of ICML, NeurIPS, ICLR, CVPR, and WACV. He is an Associate Editor of IEEE Transaction on Image Processing and IEEE Transaction on Biometrics, Behavior, and Identity Science.