HARD-Net: Hardness-AwaRe Discrimination Network for 3D Early Activity Prediction

Tianjiao Li[1,2], Jun Liu[2,⋆], Wei Zhang[1,⋆], and Lingyu Duan[3]

[1] School of CSE, Shandong University, Jinan, China
tianjiao.lee@mail.sdu.edu.cn, davidzhang@sdu.edu.cn
[2] ISTD Pillar, Singapore University of Technology and Design, Singapore
jun_liu@sutd.edu.sg
[3] School of EE & CS, Peking University, Beijing, China
lingyu@pku.edu.cn

Abstract. Predicting the class label from the partially observed activity sequence is a very *hard* task, as the observed early segments of different activities can be very similar. In this paper, we propose a novel Hardness-AwaRe Discrimination Network (HARD-Net) to specifically investigate the relationships between the similar activity pairs that are hard to be discriminated. Specifically, a Hard Instance-Interference Class (HI-IC) bank is designed, which dynamically records the hard similar pairs. Based on the HI-IC bank, a novel adversarial learning scheme is proposed to train our HARD-Net, which thus grants our network with the strong capability in mining subtle discrimination information for 3D early activity prediction. We evaluate our proposed HARD-Net on two public activity datasets and achieve state-of-the-art performance.

Keywords: Early Activity Prediction, Action/Gesture Understanding, 3D Skeleton Data, Hardness-Aware Learning

1 Introduction

Early human activity prediction (predicting the class label of an *action* or *gesture* before it is completely performed) is an important and hot research problem in the human behavior analysis domain, thanks to its relevance to many real-world applications, such as online human-robot interactions, self-driving vehicles, and security surveillance [10, 40, 42]. Existing works [32, 39] show that the 3D skeleton data [25, 29, 35, 3, 4, 2, 5, 45], which can be conveniently acquired with low-cost depth cameras, is a concise yet informative and powerful representation for human behavior analysis. Therefore, in this paper, we focus on the task of early human activity prediction from the 3D skeleton data, namely, 3D early activity prediction.

⋆ Corresponding authors.

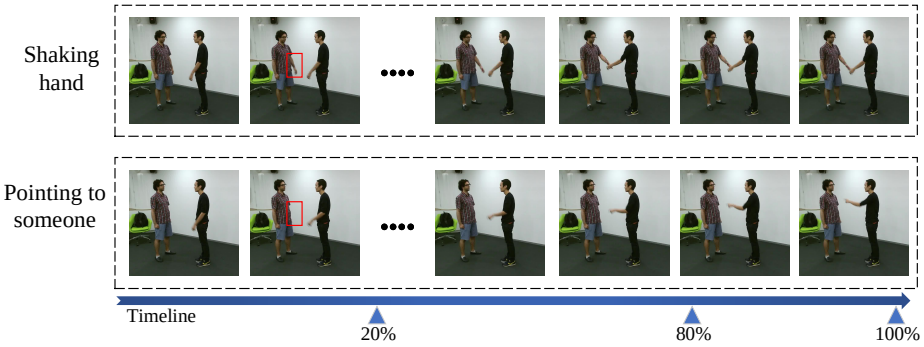

Fig. 1. Illustration of two example activities from NTU RGB+D dataset [32]. Though sufficient discrimination information can be used to distinguish these two activities when their full sequences are observed, at the early stages (e.g., when only 20% is observed), these two activities are quite similar (with only subtle discrimination information contained, as labelled by the red boxes). This makes early prediction hard.

Unlike 3D activity recognition where the full-length skeleton sequences can be used, which often contain sufficient discrimination information, in 3D early activity prediction, only the beginning segments of the sequences are observed. This makes early activity prediction much more challenging than recognition. More specifically, when performing early activity prediction, the observed beginning segments of many activities can be very similar, i.e., there may be only subtle discrepancies among them for discrimination. Thus due to the lack of significant discrimination information, these partially observed segments can be easily "mis-predicted" into other categories. For example, in Fig. 1, at the early stage (20% observation ratio), the "pointing to someone" sample can be easily mis-predicted into the "shaking hand" class, since there are only very subtle differences between them. Here we call the easily mis-predicted segments as the *hard instances*, and the classes that they are easily mis-predicted into as their *interference classes*. We also call the pair containing a *hard instance* and its corresponding *interference class* as a *hard pair*.

To deal with the challenging task of 3D early activity prediction, some existing works [13, 40] focus on inferring or distilling information from the full activity sequences that contain more sufficient discrimination information, to assist activity prediction from the partial sequences. Though remarkable progress has been achieved by the previous methods [10, 11, 42], most of them do not explicitly consider the *hard pair* discrimination issue, i.e., specifically investigating the relationships within each *hard pair* in order to exploit their minor discrepancies for better early activity prediction.

As mentioned above, the high similarities of the partially-observed activity sequences between the *hard instance* and its corresponding *interference class* make 3D early activity prediction challenging. Thus to achieve reliable prediction performance, a desired prediction model should be discriminative and powerful

enough in comprehending the relationships within the confusing *hard pair* samples and meanwhile prudentially investigating their inherent subtle discrepancies that can be exploited for discrimination.

Inspired by this, in this paper, we propose a novel Hardness-AwaRe Discrimination Network (HARD-Net), that is able to explicitly mine, perceive, and exploit the relationships and also the minor discrepancies within each *hard pair*, in order to achieve a discriminative model for early activity prediction. Concretely, in our HARD-Net, a Hard Instance-Interference Class (HI-IC) bank is specifically designed, that is able to dynamically record the *hard pairs* during the model learning procedure. Based on our HI-IC bank, an effective adversarial learning scheme for discriminating the features of the *hard pair* samples is proposed. To investigate the relationship between a *hard instance* and its *interference class*, a feature generator is designed, which produces confusing yet plausible *hard instance* features by conditioning on the similarities of this instance to the corresponding *interference class*. Meanwhile, to obtain the ability of mining subtle discrimination information within the features of *hard pair* samples, a class discriminator is further designed that pushes the prediction model to distinguish the confusing features of the *hard instance* from its *interference classes*. Therefore, with the adversarial learning going on, the generated features of the *hard instance* become more confusing with regard to its *interference class*, which in turn promote the capability of the class discriminator in mining the subtle differences that exist within the features of the *hard pair* samples for class discrimination. As a result, the proposed HARD-Net with the class discriminator as the classifier becomes very powerful in handling *hard pairs* that are often very hard to be discriminated by the early activity prediction models.

2 Related Work

3D Human Activity Recognition. Some of the existing methods [32, 26, 24, 12, 1, 46] used RNN/LSTM-based methods for 3D human activity recognition. Besides the RNN/LSTM models, 2D convolutional neural networks (CNNs) have also been investigated in this domain [14]. More recently, graph convolutional networks (GCNs) become prevalent for handling 3D activity recognition [34, 44, 38, 21, 37, 33]. Yan *et al.* [44] proposed to use spatial-temporal GCN for 3D activity recognition. Shi *et al.* [34] proposed an adaptive graph convolutional network to adaptively learn the topology of the graph for various layers, and employed the second-order information of the raw skeleton data as an extra input stream to boost the performance.

Early Human Activity Prediction. Unlike the activity recognition that is able to observe the full-length activity sequences which often contain sufficient discrimination information, in early activity prediction, only the segments from the beginning parts of the activity sequences can be used. Due to the drop of discrimination information, early activity prediction becomes much more challenging than activity recognition. Different approaches [10, 18, 13, 40, 42, 17, 19, 43, 16, 7, 23] have been proposed for early activity prediction. Ke *et al.* [13] pro-

posed to learn latent global information from full-length sequences and local information from partial-length sequences. Wang *et al.* [40] introduced a teacher-student learning architecture to transfer knowledge from the long-term sequences to the shorter-term sequences.

Overall, the aforementioned works on 3D early activity prediction do not focus on improving the discrimination ability of the prediction model by specifically handling the very similar *hard pair* samples, though the discrimination ability for the *hard pairs* is often one of the bottlenecks in early activity prediction. Different from these works, we construct an HI-IC bank to explicitly and dynamically record the *hard pair* samples, and propose a novel HARD-Net with adversarial learning to push the prediction model to be able to specifically discriminate *hard pair* samples by exploiting their relationships and mining their subtle discrimination information.

Hard Example Learning. Explicitly learning from hard-to-predict examples has been shown to be very helpful for a wide range of computer vision and machine learning tasks [27, 36, 22, 9, 28, 41, 6]. For example, Shrivastava *et al.* [36] proposed a hard example mining scheme to automatically select hard data to improve the object classification performance. Felzenszwalb *et al.* [6] proposed a margin-sensitive method for handling hard negative examples with a latent SVM to iteratively fix the latent values for positive examples and optimize the latent SVM objective function.

Unlike these methods on hard example learning, we focus on improving the ability of mining subtle discrimination information within the *hard pairs*, by explicitly pairing the easily mis-predicted early activity segments with their corresponding *interference classes* via an adversarial learning scheme. An HI-IC bank is also introduced to specifically store the *hard instances* and their *interference classes*, in order to facilitate the comprehending of the relationships and minor differences within the pairs. This thus boosts the discrimination capability of the early activity prediction model.

3 Method

3.1 Problem Formulation

Given a full-length activity sequence $S = \{s_t\}_{t=1}^{T}$, where s_t denotes the t_{th} frame, and T represents the sequence length, following existing works [10, 13], the full-length sequence S is first divided into N segments, i.e., each segment contains $\frac{T}{N}$ frames. Thus a partial sequence can be denoted as $P = \{s_t\}_{t=1}^{\tau}$, where $\tau = i \cdot \frac{T}{N}$ and $i \leq N$. The task of early activity prediction is to identify the activity category $c \in \mathbb{C} = \{1, 2, ..., C\}$ that the partial activity sequence P belongs to, based on various observation ratios.

3.2 Hardness-AwaRe Discrimination Network

3.2.1 Overview. The overall architecture of our end-to-end Hardness-AwaRe Discrimination Network (HARD-Net) is shown in Fig. 2. As mentioned above,

certain activities can be quite similar at their early stages. Thus 3D early activity prediction often suffers from lack of sufficient discrimination information when at low observation ratios. Here we introduce a new method that is able to explicitly record the hard pairs, that lack sufficient discrimination information, using an HI-IC bank, and investigate the relationship between the *hard instances* and their corresponding *interference classes* via an adversarial learning scheme, which thus enhances the capability of our prediction model in mining the minor discrimination information within the *hard samples* in feature space, for better activity prediction.

3.2.2 Hard Instance-Interference Class (HI-IC) Bank. We design an HI-IC bank to record *hard pairs*, where each pair contains a *hard instance* as well as its corresponding *interference class*, as shown in the top part of Fig. 2. This thus enables our model to get aware of the specific categories that a hard partial activity sequence can be easily mis-predicted into.

As shown in Fig. 2, the base structure of our network includes a feature encoder $\mathbb{E}$ that learns features for the partial activity sequence, and a classifier (denoted as class discriminator in Fig. 2) for class prediction. At each training iteration, each partial activity sequence (P) is fed to the encoder to obtain the original features f^{ori}, which are then further fed to the class discriminator to produce the prediction scores $\hat{y}$. If the partial sequence instance (P) is wrongly predicted by the class discriminator, given the prediction scores $\hat{y}$, the activity class c_{r_1} that has the rank-1 score in $\hat{y}$ is considered as the *interference class* (c^I) of P (i.e., $c^I = c_{r_1}$), as c_{r_1} has the most ambiguous information regarding to P.

We regard the wrongly predicted partial sequence (P) that does not have sufficient discrimination information, as the *hard instance* (P^H). We can then pack the pair of P^H and c^I as a *hard pair* that is then stored into the HI-IC bank, as illustrated in Fig. 2.

3.2.3 Adversarial Hardness-AwaRe Discrimination Learning Scheme. To investigate the relationship within the *hard pair*, we design a feature generator ($\mathbb{G}$) conditioning on the *hard instance* and its corresponding *interference class* from the pair, in order to derive latent features that are confusing yet plausible for representing the original *hard instance*. Meanwhile, to enhance the capability of our network in mining subtle discrimination information, we design a class discriminator ($\mathbb{D}^{cls}$) by granting the prediction model with the power of distinguishing the generated latent features of the *hard instance* from its *interference class*. With such an adversarial learning scheme, the generated latent features of the *hard instance* become more and more confusing with regard to its *interference class*, which in turn boosts the power of the class discriminator in mining the subtle discrimination information to distinguish the confusing *hard instance* from its *interference class*. As a result, the overall discrimination capability of the prediction model is strengthened during the adversarial learning procedure.

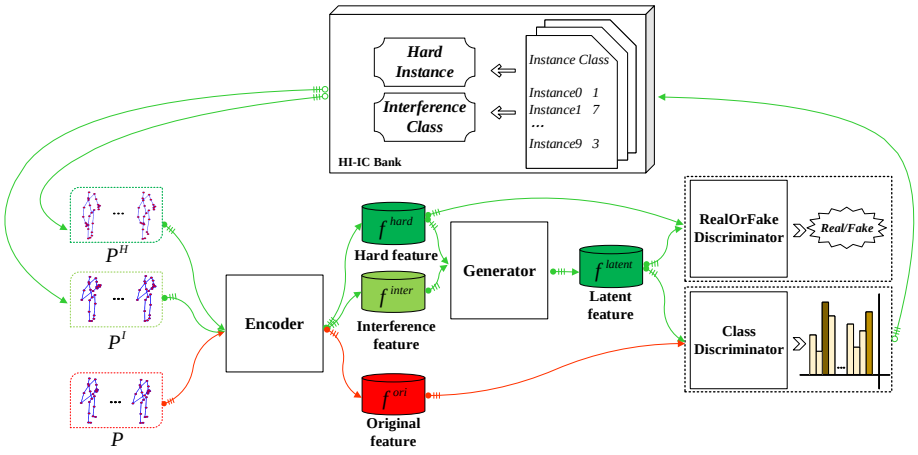

Fig. 2. Illustration of our end-to-end HARD-Net. Our network is constructed on a replaceable feature encoder (e.g., CNN skeleton encoder [13] or GCN skeleton encoder [34]) that encodes features for partial sequences. Following the red arrows representing both backbone training phase and inference phase, partial sequences (P) are fed to the encoder followed by the classifier to obtain classification scores that serve as the criterion for storing *hard pairs* into the HI-IC bank. Then following the green arrows representing adversarial learning phase, we randomly select a *hard pair* including a *hard instance* (P^H) and an *interference class* sample (P^I) from the HI-IC bank for feature encoding. The obtained feature pairs, f^{hard} and f^{inter}, are further taken into account for adversarial learning, in order to improve the capability of our prediction model in mining subtle discrepancies within each *hard pair* in the feature space.

Feature Generator. We design a generator ($\mathbb{G}$) that exploits the relationship between the *hard instance* and the corresponding *interference class*, in order to produce latent features that are confusing and hard to predict, yet are still plausible and retain inherent information representing the original *hard instance*.

Concretely, to exploit the relationship between the *hard instances* and the corresponding *interference classes*, for a *hard instance* (P^H), we refer to the HI-IC bank and identify its *interference class*. We then randomly sample an interference instance (P^I) from the *interference class*. Thus we get the paired hard samples (P^H and P^I). These two samples are fed to the feature encoder to obtain the feature pair (f^{hard} and f^{inter}) representing these two samples.

In our design, we aim to generate latent features (f^{latent}) of the *hard instance* that are very confusing and ambiguous w.r.t its *interference class*. Thus beside feeding the original features (f^{hard}) of the *hard instance* to the generator ($\mathbb{G}$), the *interference class* features (f^{inter}) are used as the interference information and are also fed to $\mathbb{G}$, as shown in Fig. 2. Therefore, conditioning on the aggregated features of P^H and P^I, the produced latent features (f^{latent}) from $\mathbb{G}$ become very confusing and hard for discriminating between these two classes.

Moreover, to specifically ensure that f^{latent} are ambiguous and hard enough for activity prediction, we further introduce an "ambiguous label" for the *hard instance* to assist the learning of $\mathbb{G}$. This "ambiguous label" can be explained as follows. Usually, in classification, the ground-truth label of the category j is a one-hot vector (y), in which the j_{th} element is set to 1 and other places are set to 0. Unlike this one-hot label, our "ambiguous label" here is represented as a vector y^{amb}, where two positions, that correspond to the ground-truth category of the *hard instance* and the category of its *interference class*, are both set to 0.5, and all other elements are set to 0.

Such an "ambiguous label" (y^{amb}) can then be used as the constraint to drive the generated latent features f^{latent} to be ambiguous between these two classes. This constraint can be formulated as follows:

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

where K denotes the total number of the activity classes, and $\hat{y}^{latent}$ is the output vector of the class discriminator that performs classification based on the generated latent features (f^{latent}).

The constraint in Eq. (1) ensures that the generated latent features (f^{latent}) are ambiguous enough. However, as mentioned before, f^{latent} still needs to be plausible and retain inherent information for representing the original *hard instance*. To achieve this, we apply a real-or-fake constraint on f^{latent} to make it plausible, as well as a mean absolute error constraint to drive f^{latent} to be closer to the features (f^{hard}) of *hard instance*.

The mean absolute error constraint ($\mathcal{L}_{con}^{\mathbb{G}}$), for narrowing the distance between f^{latent} and f^{hard}, is formulated in Eq. (2). The real-or-fake constraint ($\mathcal{L}_{rof}^{\mathbb{G}}$), brought by the RealOrFake Discriminator ($\mathbb{D}^{rof}$) for ensuring that the generated features (f^{latent}) and the original features (f^{hard}) still stay in the same feature domain, is formulated in Eq. (3).

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

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

The overall objective function for the generator ($\mathbb{G}$) can thus be formulated as:

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

With the above objective function, $\mathbb{G}$ is thus able to generate latent features (f^{latent}) that are very confusing with regard to the *interference class*, yet still retain inherent information for representing the input *hard instance*.

Class Discriminator. To obtain strong discrimination power, we design a class discriminator ($\mathbb{D}^{cls}$) that is able to distinguish the generated latent features (f^{latent}) of each *hard instance* from its corresponding *interference class*. To learn our class discriminator, a classification constraint ($\mathcal{L}^{\mathbb{D}^{cls}}$) is applied on $\mathbb{D}^{cls}$ that

pushes it to predict the accurate label (y) of the original *hard instance* based on the confusing latent features (f^{latent}):

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

Therefore, with the adversarial learning going on, the generated latent features (f^{latent}) for representing the *hard instance* become more and more confusing with regard to its *interference class* (i.e., contain less and less discrimination information for $\mathbb{D}^{cls}$ to do class distinguishing). This, however, in turn boosts the power of $\mathbb{D}^{cls}$ in comprehending the remaining subtle discriminative information in f^{latent} for distinguishing it from its *interference class*, i.e., $\mathbb{D}^{cls}$ thus becomes more and more powerful in mining the very minor discrimination information for class distinguishing.

Note that during adversarial learning, beside feeding in the generated latent features (f^{latent}) to train $\mathbb{D}^{cls}$, the original features (f^{ori}) encoded from the original samples are also fed to $\mathbb{D}^{cls}$ during training, as shown in Fig. 2. Therefore the below objective function is also applied when learning $\mathbb{D}^{cls}$:

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

As mentioned before, in our adversarial learning scheme, the original features and the generated features are kept in the same domain. Thus such a training scheme (combining Eq. (5) and (6)) is able to stabilize the overall network learning, which further yields a powerful $\mathbb{D}^{cls}$ for mining subtle discrimination information contained in both the latent features and the original features for class distinguishing. Therefore the obtained class discriminator $\mathbb{D}^{cls}$, that has very strong power in mining subtle discrimination information for distinguishing the *hard instances* from the *interference classes*, can act as the final classifier for activity prediction.

3.2.4 Training and Testing. Each training iteration of our HARD-Net is comprised of two phases, namely backbone training with HI-IC bank populating, and adversarial learning.

Backbone training & HI-IC bank filling. The backbone of our network mainly consists of an encoder $\mathbb{E}$ and a class discriminator $\mathbb{D}^{cls}$, as shown in Fig. 2. This backbone can be trained based on Eq. (6). To fill the HI-IC bank, a mini-batch of original partial sequences P with batch size B is first fed to the encoder $\mathbb{E}$ to extract features f^{ori}. Then based on f^{ori}, the class discriminator produces predicted scores, which serve as a criterion for storing *hard pairs* into our HI-IC bank, i.e., if a sample is mis-predicted, this sample and its mis-predicted class are packed as a *hard pair*, which will be stored into the HI-IC bank.

Adversarial learning scheme. During adversarial learning, the parameters of the encoder $\mathbb{E}$ are first frozen. We then sample rB *hard pairs* from the HI-IC

Algorithm 1: Learning procedure of our HARD-Net.

Input: Partial skeleton sequences (P) and ground-truth labels (c^τ)

while *not converge* **do**
  Backbone learning and HI-IC 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*($\hat{y}$);
      HI-IC Bank $\leftarrow \{P^H; c^I\}$;
    end
  end
  Adversarial HARD-Net Learning
    Freeze $\mathbb{E}$;
    Select and sample P^H and P^I from HI-IC Bank;
    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 $\mathbb{G}$;
  end
end

bank, where $0 < r \leq 1$. If there are not enough pairs in the bank (i.e., at the early stage of training process), all pairs in the bank are selected and repeated to reach rB. Otherwise, we follow the first-in-first-out strategy to select the rB *hard pairs*. Based on the *interference class* from each sampled *hard pair*, we sample an instance of it from the dataset as P^I. After that, P^H and P^I are fed into the encoder $\mathbb{E}$ for feature encoding, and then the encoded features f^{hard} and f^{inter} are fed into the generator $\mathbb{G}$ to attain f^{latent}. The generated latent features f^{latent} are fed into the RealOrFake discriminator $\mathbb{D}^{rof}$ and the class discriminator $\mathbb{D}^{cls}$ with ground-truth category to update $\mathbb{D}^{rof}$ and $\mathbb{D}^{cls}$. Finally, the f^{latent} are used to update $\mathbb{G}$ with the ambiguous label. This training procedure is detailed in Alg. 1.

Testing. As shown by red arrows in Fig. 2, at the inference phase, we input a partial skeleton sequence to the encoder to obtain features, which are then fed to $\mathbb{D}^{cls}$ for activity prediction. As $\mathbb{D}^{cls}$ has strong capabilities in mining subtle discrimination information for distinguishing hard samples from their similar interference classes, our network becomes powerful in early activity prediction.

4 Experiments

We test the proposed method for 3D early action prediction on the NTU RGB+D dataset [32], and 3D early gesture prediction on the First Person Hand Action (FPHA) dataset [8]. We conduct extensive experiments on these two datasets as below.

NTU RGB+D dataset [32] is a large dataset that has been widely used for 3D action recognition and 3D early action prediction. It contains more than 56 thousands videos and over 4 million frames from 60 activity categories. Each human skeleton in the dataset possesses 25 human body joints represented by 3D coordinates. This dataset is very challenging for 3D early action prediction, as it contains a large number of samples that are confusing at the beginning of the activity sequences. There are two standard evaluation protocols provided by the dataset. The first protocol is the Cross Subject (CS) protocol, where 20 subjects are employed for training and the remaining 20 subjects are left for testing. The second protocol is the Cross View (CV) protocol, where two viewpoints are employed for training, and the third one is for testing.

First Person Hand Action (FPHA) dataset [8] is a challenging 3D hand gesture dataset. The samples in this dataset are the first-person hand activities interacting with 3D objects recorded by six subjects. It contains over 100K frames of 45 different hand activity categories. Each hand skeleton attains 21 hand joints interpreted by 3D coordinates. We test our method on FPHA by following the standard evaluation protocol as [8], where 600 activity sequences are employed for training and the remaining 575 activity sequences are for testing.

Evaluated Models. To test the efficacy of our method, we test two different models, namely "w/o HARD-Net" and "w/ HARD-Net". (1) "w/o HARD-Net": This is actually the backbone model of our network, that contains the feature encoder and the classifier. (2) "w/ HARD-Net": This is our proposed activity prediction model (HARD-Net) that has strong capabilities in discriminating *hard pair* samples via Hardness-AwaRe Discrimination adversary leaning.

4.1 Implementation Details

To comprehensively evaluate the efficacy of our HARD-Net, we specifically construct our method above two state-of-the-art baseline encoders, namely the CNN encoder [20] and the GCN encoder [34], as shown in Tab. 4. The details of these two baseline encoders can be found in the corresponding papers [20, 34]. We also design our generator and real-or-fake discriminator by following Radford *et al.* [31], and implement the class discriminator based on multi-layer perceptron. The weights λ_1 and λ_2 in Eq. (4) are both set to 1.

All experiments are performed based on the Pytorch framework. Adam [15] algorithm is used to train our end-to-end network. The batch size B, learning rate, betas and weight decay are set to 128, 2×10^{-4}, (0.9, 0.999), and 1×10^{-5}, respectively. We set the size of HI-IC bank to be 5000 for the very large NTU RGB+D dataset and 100 for the much smaller FPHA dataset. In each training

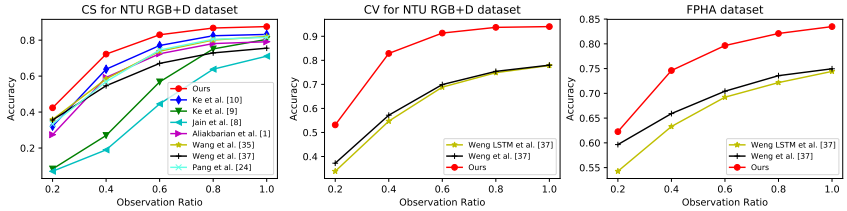

Fig. 3. Comparison of 3D early activity prediction performance on NTU RGB+D and FPHA datasets. Our method outperforms state-of-the-arts by a large margin.

Table 1. Performance comparison (%) on NTU RGB+D (cross-subject protocol). Our method outperforms the backbone model ("w/o HARD-Net") significantly. It also outperforms the state-of-the-art 3D early activity prediction methods by a large margin.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
Ke *et al.* [14]	8.34	26.97	56.78	75.13	80.43	45.63
Jain *et al.* [12]	7.07	18.98	44.55	63.84	71.09	37.38
Aliakbarian *et al.* [1]	27.41	59.26	72.43	78.10	79.09	59.98
Wang *et al.* [40]	35.85	58.45	73.86	80.06	82.01	60.97
Pang *et al.* [30]	33.30	56.94	74.50	80.51	81.54	61.07
Weng *et al.* [42]	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
w/o HARD-Net	37.82	67.87	79.22	83.39	84.52	66.91
w/ HARD-Net	**42.39**	**72.24**	**82.99**	**86.75**	**87.54**	**70.56**

iteration, the proportion r between the original instances and the *hard pair* instances used for network learning is set to $4:1$.

4.2 Experiments on 3D Early Action Prediction

We compare the proposed HARD-Net with the state-of-the-art approaches on NTU RGB+D. The comparison results with different observation ratios are shown in Tab. 1 (cross subject protocol), and Tab. 2 (cross view protocol). **Results on Cross Subject Protocol.** Comparison results on cross subject protocol are shown in Tab. 1 and Fig. 3 (left). As shown in Tab. 1, our proposed HARD-Net achieves the best performance over all observation ratios, which indicates the efficacy of our proposed HARD-Net. Compared to the state-of-the-art works and the backbone model, our method outperforms them significant, especially when the observation ratio is very low. The significant improvements demonstrate that our proposed approach can mine minor yet significant discrepancies for discrimination.

Moreover, following [30, 1, 42], we also use the area under curve metric, denoted as AUC, which is used to illustrate the average precision over all obser-

Table 2. Performance comparison (%) on NTU RGB+D (cross-view protocol). We observe only [42] has reported early action prediction results on the cross-view protocol.

	Observation Ratios					
Methods	20%	40%	60%	80%	100%	AUC
LSTM [42]	33.86	54.70	68.85	74.86	77.84	57.93
Weng *et al.* [42]	37.22	57.18	69.92	75.41	77.99	59.71
w/o HARD-Net	47.71	78.95	88.49	91.51	91.79	75.50
w/ HARD-Net	**53.15**	**82.87**	**91.34**	**93.71**	**94.03**	**78.84**

vation ratios to investigate the overall efficacy of our proposed HARD-Net. As shown in Tab. 1, our approach achieves the highest AUC of 70.56%, compared to the existing methods and also the backbone model ("w/o HARD-Net"). Note that our HARD-Net outperforms the backbone model by 3.65%, which further demonstrates that the proposed adversarial learning scheme can well-perceive and comprehend the subtle differences within *hard classes* and facilitate the discrimination capabilities of the class discriminator.

Ablation study on different loss weights in Eq. 4 are also conducted. Our method achieves AUC 70.6% under full losses ($\lambda_1 = \lambda_2 = 1$). Below we analyze the impact of each loss: 1)When removing ambiguous loss (setting its weight to 0) and keeping other two losses, the AUC drops to 67.9%. 2)When removing reconstruction loss and keeping other two losses, AUC drops to 67.7%. 3)When removing real/fake loss and keeping other two losses, AUC drops to 68.0%.

Results on Cross View Protocol. We also evaluate our HARD-Net on cross view protocol as in [42] and the comparison results are shown in Tab. 2 and Fig. 3 (middle). As shown in Tab. 2, our proposed HARD-Net model outperforms the existing works by a large margin over all observations ratios, which demonstrates the efficacy of our approach.

It is worth noting that the average accuracy score AUC of the HARD-Net exceeds the previous work [42] by 19.13% and exceeds baseline encoder by 3.34% which indicates our class discriminator can benefit from adversarial learning scheme and obtain more discrimination abilities for 3D early activity prediction.

4.3 Experiments on 3D Early Gesture Prediction

To demonstrate the efficacy of our proposed HARD-Net on 3D gesture dataset, extensive experiments are conducted on a publicly available 3D hand gesture dataset, namely FPHA. As illustrated in Tab. 3 and Fig. 3 (right), our proposed HARD-Net achieves better performance consistently over all observation ratios compared to Weng *et al.* [42].

Compared to the baseline model, at the early stages when the observation ratio is very low that lack sufficient discrimination information, since our HARD-Net is powerful in mining minor discrepancies, it achieves the most significant performance gain by 9.57% at the 20% observation ratio.

Table 3. Quantitative results (%) comparison on FPHA with state-of-the-arts.

Methods	Observation Ratios					
	20%	40%	60%	80%	100%	AUC
LSTM [42]	54.26	63.30	69.22	72.17	74.43	64.11
Weng *et al.* [42]	59.65	65.91	70.43	73.57	74.96	66.66
w/o HARD-Net	62.26	74.61	79.65	82.09	83.48	72.17
w/ HARD-Net	**71.83**	**82.78**	**86.09**	**87.13**	**87.30**	**78.56**

4.4 Ablation Study

In this section, extensive ablation experiments are conducted based on the NTU-RGB+D dataset (cross-subject protocol), that is widely used by existing works [13, 30, 40, 42] in early activity prediction community.

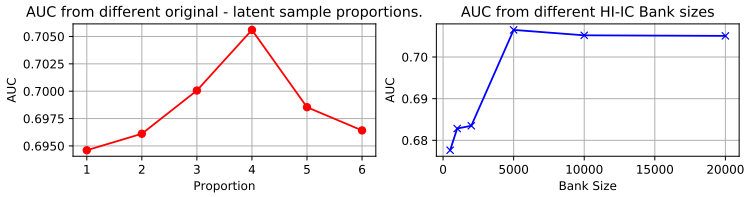

Fig. 4. Left: Evaluation of the impact of using different proportions of original samples and *hard pair* samples for network training. When proportion between original sample number and *hard pair* sample number is 4:1, our model achieves the highest prediction accuracy. Right: Evaluation of the impact of different HI-IC bank sizes.

Impact of Bank Size. We evaluate the performance of the HI-IC bank in different bank sizes. The result is shown in Fig. 4 (right). The AUC reflecting average precision over all observation ratios increases rapidly from a smaller bank size to a larger one, and then remains stable when the bank size is large enough (e.g., size 5000). This can be explained by the number of *hard pairs* in a dataset, and when the intrinsic threshold is reached, the further performance gain is limited.

Impact of proportions between original features and latent features for training. Our experimental results in Fig. 4 (left) show that the optimal proportion of original features and latent features used for network training is $4:1$. This can be explained as: if we use too much original features for network training, then less useful discrimination information will be mined via our adversarial learning scheme. However, if too much latent features are employed for training, it may lead to performance drops over the original samples. Moreover, small performance differences achieved by different ratios (1:1 to 6:1) indicate

Table 4. Performance gain (%) brought by our HARD-Net with different backbones.

Backbone	Methods	Observation Ratios				
		20%	40%	60%	80%	100%
CNN backbone [13]	w/o HARD-Net	34.01	63.16	75.87	81.39	82.24
	w/ HARD-Net	**35.86**	**64.97**	**77.12**	**82.22**	**82.98**
	Δ	**+1.85**	**+1.81**	**+1.25**	**+0.83**	**+0.74**
GCN backbone [34]	w/o HARD-Net	37.82	67.87	79.22	83.39	84.52
	w/ HARD-Net	**42.39**	**72.24**	**82.99**	**86.75**	**87.54**
	Δ	**+4.57**	**+4.37**	**+3.77**	**+3.36**	**+3.02**

that our HARD-Net is not sensitive to ratios. In Fig. 4 (left), all achieved AUCs of the HARD-Net are in a small range (69.5% to 70.6%), which shows robustness of our method against ratios. Besides, these AUCs achieved all outperform the baseline (66.9%) by a large margin, validating the efficacy of our HARD-Net.

Impact of Backbone Encoder. We extensively test our algorithm on a CNN backbone and a GCN backbone, and show the efficacy of the proposed method. As shown in Tab. 4, our HARD-Net boosts early prediction performance on both backbone models obviously, especially at the very low observation ratios. This indicates that our HARD-Net is powerful in mining subtle discrimination information for early activity prediction.

5 Conclusion

In this paper, we propose a novel Hardness-AwaRe Discrimination Network (HARD-Net) for 3D early activity prediction. The proposed HARD-Net is able to explicitly investigate the relationship between an easily mis-predicted instance, named *hard instance*, and the particular category that it is mis-predicted into, named *interference class*. An adversarial learning scheme is proposed to mine subtle discrepancies between this *hard instance* - *interference class* pair by generating ambiguous and less discriminative latent features conditioned on that particular pair to represent original *hard instances*. We further design a class discriminator to distinguish the derived latent features from the corresponding *interference classes*. With such a network design, our proposed HARD-Net achieves state-of-the-art performance on two challenging datasets.

Acknowledgement. This work is supported by SUTD Project PIE-SGP-Al-2020-02, SUTD Project SRG-ISTD-2020-153, the National Natural Science Foundation of China under Grant 61991411, and Grant U1913204, the National Key Research and Development Plan of China under Grant 2017YFB1300205, and the Shandong Major Scientific and Technological Innovation Project (MSTIP) under Grant 2018CXGC1503.

References

1. Aliakbarian, M., Saleh, F., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Encouraging lstms to anticipate actions very early (2017)
2. Cai, Y., Ge, L., Cai, J., Magnenat-Thalmann, N., Yuan, J.: 3d hand pose estimation using synthetic data and weakly labeled rgb images. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
3. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular rgb images. In: Proceedings of the European Conference on Computer Vision. pp. 666–682 (2018)
4. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2272–2281 (2019)
5. Cai, Y., Huang, L., Wang, Y., et al.: Learning progressive joint propagation for human motion prediction. In: Proceedings of the European Conference on Computer Vision (2020)
6. Felzenszwalb, P., Girshick, R., Mcallester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**, 1627–45 (09 2010)
7. Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Predicting the future: A jointly learnt model for action anticipation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5561–5570 (2019)
8. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018)
9. Girshick, R.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1440–1448 (Dec 2015)
10. Hu, J.F., Zheng, W.S., Ma, L., Wang, G., Lai, J.: Real-time rgb-d activity prediction by soft regression. In: European Conference on Computer Vision. pp. 280–296 (2016)
11. Hu, J.F., Zheng, W.S., Ma, L., Wang, G., Lai, J., Zhang, J.: Early action prediction by soft regression. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(11), 2568–2583 (2018)
12. Jain, A., Singh, A., Koppula, H., Soh, S., Saxena, A.: Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. Arxiv (2015)
13. Ke, Q., Bennamoun, M., Rahmani, H., An, S., Sohel, F., Boussaid, F.: Learning latent global network for skeleton-based action prediction. IEEE Transactions on Image Processing **29**, 959–970 (2020)
14. Ke, Q., Bennamoun, M., An, S., Sohel, F.A., Boussaïd, F.: A new representation of skeleton sequences for 3d action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4570–4579 (2017)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
16. Kong, Y., Tao, Z., Fu, Y.: Adversarial action prediction networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(3), 539–553 (2020)
17. Kong, Y., Gao, S., Sun, B., Fu, Y.: Action prediction from videos via memorizing hard-to-predict samples. In: AAAI (2018)
18. Kong, Y., Kit, D., Fu, Y.: A discriminative model with multiple temporal scales for action prediction. In: European Conference on Computer Vision. vol. 8693 (2014)

19. Kong, Y., Tao, Z., Fu, Y.: Deep sequential context networks for action prediction. IEEE Conference on Computer Vision and Pattern Recognition pp. 3662–3670 (2017)
20. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: IJCAI. pp. 786–792 (2018)
21. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
22. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (Feb 2020)
23. Liu, J., Shahroudy, A., Wang, G., Duan, L., Kot, A.C.: Skeleton-based online action prediction using scale selection network. IEEE Trans. Pattern Anal. Mach. Intell. **42**(6), 1453–1467 (2020)
24. Liu, J., Wang, G., Duan, L., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (April 2018)
25. Liu, J., Ding, H., Shahroudy, A., Duan, L.Y., Jiang, X., Wang, G., Kot, A.C.: Feature boosting network for 3d pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 494–501 (2020)
26. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 816–833. Springer International Publishing, Cham (2016)
27. Loshchilov, I., Hutter, F.: Online batch selection for faster training of neural networks (2015)
28. Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.: 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. pp. 3235–3243 (2019)
29. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2823–2832 (2017)
30. Pang, G., Wang, X., Hu, J.F., Zhang, Q., Zheng, W.S.: Dbdnet: Learning bi-directional dynamics for early action prediction. In: IJCAI. pp. 897–903 (2019)
31. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016), http://arxiv.org/abs/1511.06434
32. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
33. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
34. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)
35. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth

images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1297–1304 (2011)
36. Shrivastava, A., Mulam, H., Girshick, R.: Training region-based object detectors with online hard example mining (2016)
37. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
38. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
39. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3d human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(5), 914–927 (2013)
40. Wang, X., Hu, J., Lai, J., Zhang, J., Zheng, W.: Progressive teacher-student learning for early action prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3551–3560 (2019)
41. Wang, X., Shrivastava, A., Gupta, A.: A-fast-rcnn: Hard positive generation via adversary for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
42. Weng, J., Jiang, X., Zheng, W., Yuan, J.: Early action recognition with category exclusion using policy-based reinforcement learning. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2020)
43. Xu, W., Yu, J., Miao, Z., Wan, L., Ji, Q.: Prediction-cgan: Human action prediction with conditional generative adversarial networks. Proceedings of the ACM International Conference on Multimedia (2019)
44. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
45. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In: Proceedings of the European Conference on Computer Vision (2020)
46. Zhu, W., Lan, C., Xing, J., Li, Y., Shen, L., Zeng, W., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI (2016)