Computing and Informatics, Vol. 44, 2025, 549–579, doi: 10.31577/cai_2025_3_549

SELF-SUPERVISED LEARNING FOR 3D ACTION PREDICTION WITH GRAPH CONVOLUTIONAL RECURRENT NETWORK

Peng Liu[§], Yifan Wang[§], Qicong Wang, Chong Zhao[*]

Department of Computer Science and Technology
Xiamen University, Xiamen 361000, China
e-mail: {delom, 23020201153803}@stu.xmu.edu.cn,
{qcwang, zhc}@xmu.edu.cn

Yan Chen[*]

College of Business and Management
Xiamen Huaxia University, Xiamen 361024, China
e-mail: chenyan@hxxy.edu.cn

Man Qi

School of Engineering, Technology and Design
Canterbury Christ Church University, Canterbury, CT1 1QU, UK
e-mail: man.qi@canterbury.ac.uk

Abstract. In view of the dependence of existing 3D action prediction research on labels, we propose a graph convolutional recurrent 3D action prediction method based on state discrimination and spatio-temporal self-supervised contrast learning. In the state discrimination task, cross-sample sampling and relative action completeness perception are used to train the model for generalized state information learning across instances and classes. In the spatio-temporal contrast task, spatio-temporal consistency information is introduced into the feature representation to

[§] Co-first authors
[*] Corresponding authors

enrich action semantics in features. Additionally, in order to fully extract spatio-temporal information in 3D action sequences, a spatio-temporal feature extraction network (STFEN) based on graph convolution recurrent network is designed. The experimental results on public datasets demonstrate the efficiency of the proposed methods.

Keywords: 3D action prediction, self-supervised learning, state discrimination, spatio-temporal consistency, contrast learning

Mathematics Subject Classification 2010: 68T01

1 INTRODUCTION

With the development of information technology and the explosive growth of data in society, data-driven artificial intelligence has shown increasingly powerful perception, understanding, judgment, prediction, and even creation capabilities, and is widely used in various fields such as image recognition, video understanding, speech recognition, machine translation, autonomous driving, and recommendation systems, which greatly facilitate people's lives and improve productivity. The current mainstream artificial intelligence methods are implemented based on deep learning techniques [1, 2], by building deep neural network models to extract features from the input data, and by training the models to adjust their own parameters through a large amount of data, so that the models can acquire certain wisdom capabilities in the process of learning.

Action understanding is an important direction for AI applications, which aims to understand the inner action semantics of human actions based on their external performance. Depending on the input data, the action understanding methods can be divided into video-based methods [3, 4, 5, 6, 7] and skeleton-based 3D methods. Unlike video data, 3D human skeleton data, which can be easily obtained from sensors [8], has a more direct representation capability by excluding the interference of scene information and extraneous objects. Therefore, the method in this paper takes the 3D skeleton data as input. Self-supervised learning trains models to learn feature representations without manual labeling by designing agent tasks that can automatically generate labels based on the characteristics of the data itself. Because self-supervised learning can reduce the dependence of models on expensive labels, many researchers have used self-supervised methods to investigate 3D action understanding in recent years, and most of them have focused on 3D action recognition [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25].

Unlike 3D action recognition, which takes a complete action sequence as input, 3D action prediction takes an incomplete skeleton sequence as input and outputs its action category, which has broader application prospects in the fields of intelligent security, human-computer interaction, autonomous driving, health monitoring, etc. This requires the features extracted by the model to be semantically general and distinguishable in order to ensure the accuracy of the prediction. There are relatively few studies on 3D action prediction, and they all focus on the supervised learning paradigm [26, 27, 28, 29, 30, 31, 32, 33, 34, 35]. In order to utilize the unlabeled data, this paper proposes to train the 3D action prediction model by self-supervised learning method.

Specifically, we design two self-supervised tasks to train the model to extract state information and spatio-temporal consistency information in incomplete sequences. In an incomplete action sequence, the human body has performed the action to a certain extent, which indicates the state of action execution. In the action prediction task, this state can be represented by the observation rate. For example, a sequence with an observation rate of 0.3 indicates that 30% of the action in this sequence has been performed. In this paper, we let the model perceive the state of action completeness, i.e., the observation rate of an incomplete action sequence, and compare it with the ground truth observation rate to produce a loss. However, perceiving the absolute observation rate of action from a single sample only will lead the model to fit only on samples from a single instance during the learning process and ignore the connection between data, and the learning of action state knowledge is limited to an action category, which cannot effectively learn the common features of action states that are widely present in all kinds of samples. According to the psychological study [36], humans are better at making relative judgments among different information than making absolute judgments about a single piece of information. Inspired by the above factors, the state discrimination task is proposed in this paper. This self-supervised task converts the absolute action completeness perception task into a relative action completeness perception task, and obtains state discriminant pairs based on cross-sample sampling, thus extending the learning scope of the model for action state representation and introducing more generalizable cross-category state information into the feature representation.

With the progress of self-supervised learning research, contrast learning [37, 38] has received increasing attention from researchers, and many self-supervised methods [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24] learn feature representations by contrast learning on 3D action recognition tasks. In the process of human movement, the spatial properties of human skeleton joints are important carriers of action semantics, and incomplete action sequences have a diversity of time-domain distribution, so spatio-temporal features are important basis for 3D action prediction by the model. In this paper, a novel cross spatio-temporal contrast learning framework is proposed to learn the feature representation with temporal invariance by transforming the samples from both the time domain and the space domain for the action prediction task itself. Specifically, the proposed method introduces skeleton sequence original sample sampling and spatial augmentation, and performs temporal contrast, spatio contrast, and spatio-temporal cross-contrast learning on the basis of obtaining temporal transformed samples, spatio transformed samples, and spatio-temporal transformed samples, respectively, to train the model to learn the feature representation with temporal consistency, spatial consistency, and spatio-temporal consistency.

In addition, existing action prediction methods either use RNN as the feature encoder [31], for each frame of the human skeleton joint point coordinate set, which is linearly arranged and input to the network to extract features, this method does not take into account the natural connection structure of the human skeleton, resulting in the model cannot effectively perceive the spatial information of human motion; or use GNN as the feature encoder [32, 33, 34, 35], this method builds a graph of the human skeleton, which has a strong ability to capture spatial features, but for temporal features is carried out through the convolution between neighbouring frames, which lacks global temporal domain information. For the action prediction task, whose input has temporal incompleteness, sensitivity to temporal information enables the model to understand the action better, and the model should also have the ability to perceive spatial features well. Therefore, a spatio-temporal feature extraction network (STFEN) based on graph convolution recurrent is proposed in this paper. The network consists of two modules: spatial domain feature extraction and temporal domain feature transfer. In the spatial feature extraction module, the graph is constructed based on the a prior knowledge of the skeleton structure of the human body, and the graph convolution is used to convolve the features between joints to obtain the spatial feature containing rich spatial information; in the temporal feature transfer module, the static spatial features of human body structure in a single frame are transferred between frames to obtain the motion feature in the temporal domain. Finally, the features are aggregated to obtain the spatio-temporal action feature of the whole sequence.

The contribution of this paper can be summarized as follows:

- Based on relative action completeness perception, we devise a self-supervised agent task which can generate state discriminant pairs and rank them by action completion, so as to extract the action state common information among various types of samples.

- A novel spatio-temporal contrast method is proposed to extract temporal consistency, spatial consistency and spatio-temporal consistency about actions by performing temporal transformation, spatial enhancement and cross-contrastive learning.

- We propose a spatio-temporal feature extraction network (STFEN) based on graph convolution recurrent that hierachically extracts spatial features, temporal features, and global spatio-temporal features of 3D action sequences while preserving sensitivity to temporal information.

2 RELATED WORK

2.1 Self-Supervised 3D Action Understanding

Compared with video data, 3D action understanding based on skeleton data has received a lot of attention from researchers in recent years. Zheng et al. [9] introduced a novel conditional skeleton inpainting network to capture the long-term global motion dynamics in sequences with varying length and designed an additional adversarial training strategy which can enhance the encoder-decoder model for learning more discriminative representations. This is the first work to explore self-supervised representation learning approaches for skeleton-based action recognition. In [10], Lin et al. designed three tasks, including a generation task, a classification task, and a comparison learning to learn comprehensive and general feature representations. To utilize self-supervised learning for semi-supervised learning, Si et al. [11] proposed an adversarial self-supervised learning framework network that tightly couples self-supervised and the semi-supervised scheme via neighbour relation exploration and adversarial learning. In [25], Yang et al. formulate the unsupervised action recognition learning as an attention prediction problem, where the encoder captures action-specific motion patterns by predicting multiple self-attentions in spatio-temporal dimensions. Through contrastive representation learning by adequate compositions of viewpoints and distances, Gao et al. [12] proposed a self-supervised method to select discriminative features which have invariance motion semantics for action recognition. To learn semantic information, Xu et al. [13] designed a framework which not only creates reverse sequential prediction to learn low-level information and high-level pattern, but also devises action prototypes to implicitly encode semantic similarity shared among sequences. In [14], Rao et al. proposed a generic unsupervised contrastive action learning paradigm named AS-CAL which could perform contrastive learning on action patterns of augmented skeleton sequences, to enable the model to learn effective action representations from unlabeled skeleton data. Su et al. [15] proposed a novel system which associates the sequences with actions for unsupervised skeleton-based action recognition. The system is based on an encoder-decoder recurrent neural network, where the encoder learns a separable feature representation within its hidden states formed by training the model to perform the prediction task. Li et al. [16] proposed a Cross-view Contrastive Learning framework for unsupervised 3D skeleton-based action Representation (CrosSCLR) which consists of both single-view contrastive learning (SkeletonCLR) and cross-view consistent knowledge mining (CVCKM) modules. Su et al. [17] proposed a novel self-supervised method which constructs a positive clip (speed-changed) and a negative clip (motionbroken) of the sampled action sequence, to encourage the positive pairs closer while pushing the negative pairs to force the network to learn the intrinsic dynamic motion consistency information. Wang et al. [18] proposed a novel Contrast-Reconstruction Representation Learning network (CRRL) which mainly consists of three parts: sequence reconstructor, contrastive motion learner, and information fuser. Thoker et al. [19] proposed inter-skeleton contrastive learning, which learns from multiple different input skeleton representations in a cross-contrastive manner. In [20], extreme augmentations and Energy-based Attention-guided Drop Module (EADM) were proposed to generate diverse positive samples to construct a contrastive learning framework utilizing abundant information mining for self-supervised action representation. Chen et al. [21] proposed a contrastive learning framework with a spatio-temporal skeleton mixing augmentation (SkeleMix) to complement current contrastive learning approaches by providing hard contrastive samples. Wu et al. [22] proposed a new self-supervised agent task to optimize the initialization of model parameters by training the model to compare the temporal coherence of samples. Pang et al. [23] proposed a novel Contrastive GCN-Transformer Network (ConGT) which fuses the spatial and temporal modules in a parallel way. In [24], Zhao et al. proposed a contrast learning method combined with a temporal-masking mechanism of skeleton sequences to encourage the network able to learn action representations other than feature invariance.

There exist additional studies on behavior recognition as well. Xu et al. [39] explored a semi-supervised skeleton-based action recognition method and proposed an X-invariant contrastive augmentation and representations learning framework. By learning augmentations and representations of skeleton sequences, the rotate-shear-scale invariant features are completely obtained. At the same time, they [40] believed that existing contrastive learning methods confuse the spatio-temporal information reflecting different semantic at the frame and joint levels, and designed a spatio-temporal decouple-and-squeeze contrastive learning framework to jointly compare spatio-temporal features and global features to comprehensively learn more abundant representations. Shu et al. [41] proposed a Multi-granularity Anchor-Contrastive representation Learning to address the three limitations of contrastive learning, which can learn multi-granularity representations by conducting inter- and intra-granularity pretext tasks on the learnable and structural-link skeletons. Starting from a coarse-grained perspective, Xu and Shu [42] proposed a pyramid self-attention polymerization learning framework to learn different levels representations containing abundant and complementary semantic information through multi-granularity comparison. To mitigate the potential supervisory effect of ignoring instance information, Yan et al. [43] presented a novel framework, namely Progressive Instance-aware Feature Learning, to progressively extract, reason, and predict dynamic cues of moving instances from videos for compositional action recognition.

2.2 3D Action Prediction

D action prediction, also known as early skeleton-based action recognition, has been a research direction in the field of action understanding in recent years. Hu et al. [26] proposed a soft regression-based activity prediction model and a local accumulative frame feature (LAFF) for real-time activity prediction on RGB-D sequences. Jain et al. [27] introduced a sensory-fusion architecture which jointly learns to anticipate and fuse information from multiple sensory streams. In [28], a global regularizer was introduced to learn a uniformly distributed hidden feature space and a temporal-aware cross-entropy was designed as the classification loss for early action recognition at different observation ratios. Liu et al. [29] proposed a novel window scale selection method to make the network focus on the performed part of the ongoing action and try to suppress the possible incoming interference from the previous actions at each step. Ke et al. [30] proposed a new latent global network based on adversarial learning to learn the latent global information of the partial sequences and improve action prediction. Weng et al. [31] pre-trained a classifier of all categories, and modeled the category exclusion as a mask operation on the classification probability output of the classifier. In [32], a novel Hardness-AwaRe Discrimination Network (HARD-Net) was proposed to specifically investigate the relationships between the similar activity pairs that are hard to be discriminated. Li et al. [33] designed a novel adaptive graph convolutional network with adversarial learning (AGCN-AL) that uses adversarial learning to make the features of partial sequences as similar as possible to those of complete sequences, and introduced a temporal-dependent loss to prevent the network from paying too much attention to partial sequences whose observation ratios are small. Liu et al. [34] proposed a Graph Convolutional Network with Early Attention Module (GCN-EAM), which employs a series of spatial-temporal graph convolution blocks to extract features from skeletons. In [35], a novel two-stage knowledge distillation framework was proposed to transfer prior knowledge to assist the early prediction of ongoing actions.

In summary, the current studies on 3D action prediction are still focused on the supervised paradigm, while the self-supervised studies in the field of 3D action understanding have focused on 3D action recognition and neglected 3D action prediction. It is necessary to explore a self-supervised learning method applicable to 3D action prediction.

3 METHOD

In this section, we present our self-supervised learning method details. We firstly describe the overall framework and symbols in Section 3.1, then present the proposed state discrimination task in Section 3.2 and spatio-temporal contrast task in Section 3.3, and we provide a description of our spatio-temporal feature extraction network in Section 3.4. The main symbols are summarized in Table 1.

3.1 Overall Framework

Let $X = \{x_1, x_2, x_3 \dots x_{T-1}, x_T\}$ denote a complete 3D skeleton sequence of action lasting for a total of T frames, and the set of spatial coordinates of the human skeleton joint points in frame i is x_i. Given an observation rate O, it means that the action has executed $100 * O\%$ of the complete action. For a complete action sequence sample X, when it is under the observation rate O, the incomplete 3D action sequence sample $X_O = \left\{x_1, x_2, x_3, \dots x_{\lfloor O \times T \rfloor - 1}, x_{\lfloor O \times T \rfloor}\right\}$ is obtained by intercepting its previous $\lfloor O * T \rfloor$ frames, where $\lfloor \cdot \rfloor$ is the downward rounding operation.

Notations	Definitions
X	complete 3D skeleton sequence
O	the observation rate
X_O	the partial skeleton sequence under the observation ratio O corresponding to X
E	the feature encoder
E_s	the shared feature encoder
E_m	the momentum feature encoder
F	the feature extracted from X_O
L_D	the loss of state discrimination
L_C	the loss of spatio-temporal contrast
L_O	the loss of absolute action completeness perception
L_M	the loss of motion prediction
L_{self}	total loss of the network in the self-supervised training

Table 1. Notations and definitions

All samples in the 3D action recognition dataset are traversed, and each sample is sampled at 9 different observation rates to obtain 9 incomplete skeleton sequences of different lengths from the same original complete sample. The observation rate of each incomplete skeleton sequence and the corresponding original sample serial number r are recorded.

The goal of self-supervised learning is to train a feature encoder E, capable of representing incomplete 3D skeleton sequences as discriminative action semantic features for application to a downstream task: 3D action prediction, through agent tasks that do not require manual annotation. The proposed self-supervised learning approach simultaneously uses multiple self-supervised losses to jointly guide the network to learn feature representations suitable for 3D action prediction. The overall framework is shown in Figure 1.

In the state discrimination task, according to the observation rate of the training sample X_O and the original sample serial number r, cross-sample sampling is performed in the sample bank to obtain the discriminant sample X_D which is from different original sample and under different observation rate compared with X_O. The loss is obtained by performing relative action completeness perception of X_O and X_D, which generates supervision signal to supervise the training of the feature encoder E. The spatio-temporal feature encoder E consists of two modules, the shared encoder E_s and the momentum encoder E_m, which have identical structures. In the spatio-temporal contrast task, according to the original sample sequence number r of the training sample X_O, its corresponding complete action sequence X is obtained from the sample bank, and X_O and X are spatially augmented to obtain $X_{O,aug}$ and X_{aug}, respectively, and the four samples are sent to E_s and E_m for feature encoding to obtain their respective query features and key features, then cross-contrast learning is performed between the features of the four samples to provide supervised signals for the training of E. The state discrimination task and

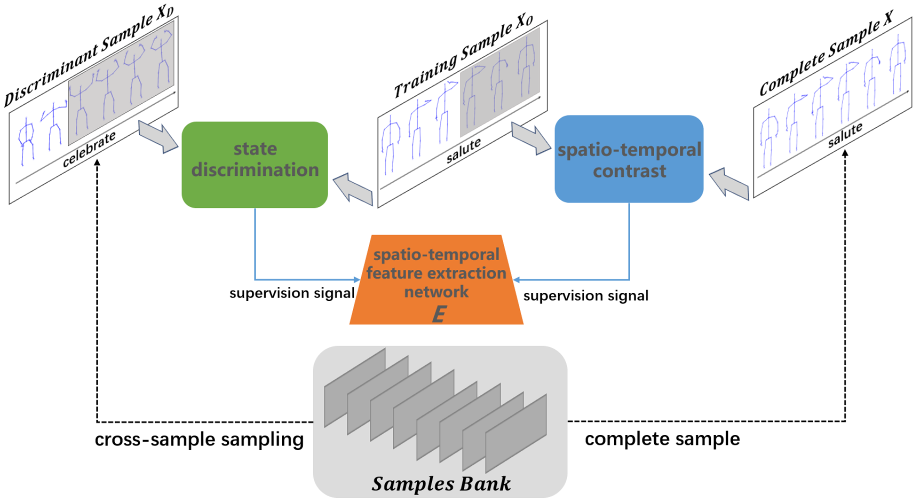


Figure 1. The overall framework of the proposed method. Our self-supervised learning framework adopts a multi-task design, including a state discrimination task and a spatio-temporal contrast task. For the training sample X_O, after obtaining its corresponding discriminant sample X_D and the complete sample X, state discrimination and spatio-temporal comparison are performed to generate supervision signals to train the proposed spatio-temporal feature extraction network, respectively.

the spatio-temporal contrast task enable the network to be trained to learn generic state information and spatio-temporal consistency information of the 3D skeleton sequence, enhancing the network's ability to represent samples.

3.2 State Discrimination

Action completeness, or observation rate O, indicates the progress of action execution at the current moment, can provide the model with a supervision signal containing information about the current state of the action. Introducing the state knowledge of the action into the feature representation can effectively enhance the model's ability to understand the observed part of an incomplete action sequence and to imagine the unobserved part, and thus improve the model's performance on 3D action prediction task.

State information about action completeness is an important element of action semantics, and learning this information can help the model understand the action better. In the action prediction task, this information is able to represent the context of the phase in which the action is being performed. Therefore, we propose the absolute action completeness perception task, for the incomplete action sequence X_O input to the model, the model perceives its action completeness and compares the perceived result with the ground truth observation rate to produce a loss. The calculation process is as follows:

$$L_O = \frac{\sum_{i=1}^{N} \|O_{P,i} - O_i\|_2^2}{N} = \frac{\sum_{i=1}^{N} \|\sigma(FC(E_s(X_{O,i}))) - O_i\|_2^2}{N}, \tag{1}$$

where L_O indicates the loss of absolute action completeness perception, N is the size of the batch and σ is the function of Sigmoid. This process is completed through a shared feature encoder E and an action completion perceptron H_O. Specifically, partial sequence is fed into E to obtain its feature representation. Subsequently, this feature representation is input to H_O to predict the observation rate of sequence. The weights of E and H_O are trained by back-propagation using this loss.

However, the use of the absolute action completeness perception task would limit the model's learning of state information to the same original sample and cannot effectively learn the common patterns about action state that coexist in various categories and instances. Moreover, according to the psychological study [36], humans are better at making relative judgments compared to absolute judgments. Based on the above premise, we further propose a state discrimination task based on relative action completeness perception to generate discriminant pairs by cross-sample sampling, which extends the learning range of the model to different original samples, and introduces cross-instance and cross-category state information into the feature representation by performing state discrimination between discriminant pairs, so that the model learns more robust action state knowledge and thus maps the 3D skeleton sequence to a feature space containing richer action semantics. The design of the state task is shown in Figure 2.

As shown in Figure 2, an incomplete action sequence X_O with observation rate O and original sample number r is given as the training sample. Firstly, cross-sample sampling is performed in the sample bank to obtain its corresponding discriminant sample. The sampling process is given by the following equation:

$$X_D = S\left[i_D\right] = S\left[\text{sample}\left(O, r\right)\right], \tag{2}$$

where S denotes the sample bank, $[\cdot]$ is the value taking operation, and $\text{sample}\left(\cdot\right)$ is the sampling function. The sampling function is calculated as follows:

$$\text{sample}\left(O, r\right) = \begin{cases} x, & \text{if } \mathit{diff}\big(x = \text{rand}\left(1, \mathit{size}\left(S\right)\right), O, r\big), \\ \text{sample}\left(O, r\right), & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathit{diff}\,(i, O, r)$ means the observation rate of the i^{th} sample in the sample bank is not equal to O and the original sample number is not equal to r. $\mathit{size}(S)$ is the size of the sample bank S.

Equation (3) indicates that in the loop, one sample from the sample bank is randomly selected each time, and if the observation rate of that sample is different from the training sample and from a different original sample, it can be used as the

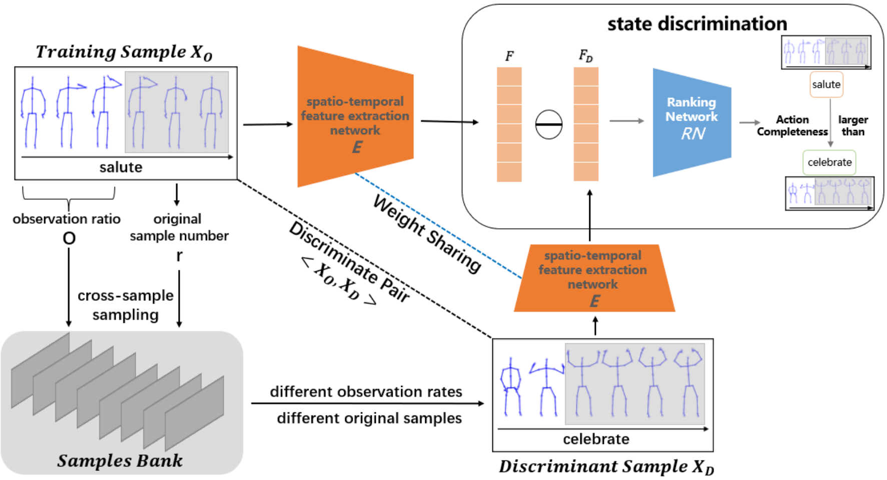


Figure 2. Flow chart of the state discrimination task. Discriminant sample pairs are constructed based on the observation rate and original sample number of the training sample. The supervised signal is generated by training the model to compare the action completeness between the two samples.

discriminant sample of X_O, otherwise the random sampling is performed again. The discriminant sample X_D is obtained by cross-sample sampling, and the discriminant pair$\langle X_O, X_D\rangle$ is constructed. Let the observation rate corresponding to X_D be O_D, and the label labelD of the state discrimination task can be obtained from the following equation:

$$\mathrm{label}_D = \begin{cases} 1, & O > O_D, \\ 0, & O < O_D. \end{cases} \tag{4}$$

As shown in Figure 2, X_O and X_D are input to the encoder E for feature extraction, and the respective features F and F_D are obtained. Then the difference between F and F_D is made to obtain the relative feature representation of the two discriminant samples. Input it into the ranking network RN to determine the relative action completeness size of the training and discriminant samples. The ranking network RN is a multi-layer perceptron capable of mapping the input to features in a high-dimensional space and downscaling to a one-dimensional output to obtain the action state ranking results *rank* for X_O and X_D. The calculation process is as follows:

$$\mathrm{rank} = RN\left(F - F_D\right). \tag{5}$$

Since label_D is a binary label, the problem can be viewed as a binary classification problem. The loss L_D for this task is obtained by applying the cross-entropy loss function:

$$\begin{aligned} L_D &= \frac{1}{N}\sum_{i=1}^{N} \mathit{CrossEntropy}(\mathrm{rank}_i, \mathrm{label}_{D,i}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(-\log\frac{\exp\left(\mathrm{rank}_i\left[\mathrm{label}_{D,i}\right]\right)}{\exp\left(\mathrm{rank}_i\left[0\right]\right)+\exp\left(\mathrm{rank}_i\left[1\right]\right)}\right), \end{aligned} \tag{6}$$

where N is the size of the batch, rank_i is the result of the ranking network to discriminate the state of the sample pair constructed from the i^{th} sample in the batch, and $\mathrm{label}_{D,i}$ is the label of the discriminant pair.

In this task, the shared encoder E_s in the spatio-temporal feature extraction network E performs feature extraction on samples, and the action discriminative loss L_D passes supervised signals through back propagation to simultaneously supervise the optimization of the parameters of the shared encoder E_s and the ranking network RN, thus introducing generalized state information across samples and categories into the feature representation of the encoder.

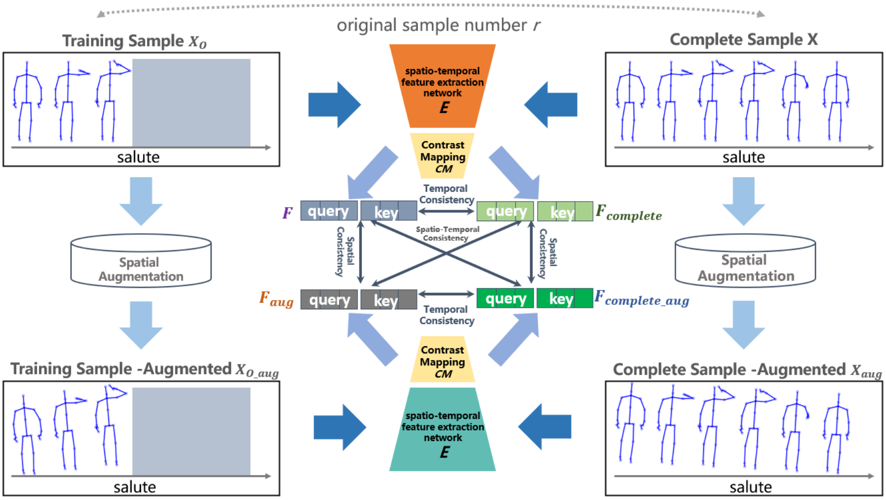


Figure 3. The framework of spatio-temporal contrast learning. For the training sample X_O, after obtaining its corresponding complete sequence X, spatial augmentation is performed on X_O and X to obtain X_{O_aug} and X_{aug}, respectively, and cross-contrast learning is performed between the four samples to learn spatio-temporal consistency information.

3.3 Spatio-Temporal Contrast

Three-dimensional human skeleton sequences contain rich temporal and spatial information. In the action prediction task, temporal information describes the current state of the action and the trend of motion change, while spatial information describes the distribution and connection of human body joints in the spatial coordinate system. Feature representations containing rich spatio-temporal information can better describe incomplete action sequences and provide a more adequate basis for action prediction.

Previous research [28] used a global regularization task to optimize the distribution of the feature space by learning the same samples under different observation rates in contrast from a temporal perspective, introducing global information from the complete sequence into the feature representation of the incomplete sequence, thus compensating to some extent for the lack of discriminative information in the incomplete sequence. However, this method only compares the samples transformed in the temporal domain, and the learned feature representations are only temporally consistent and lack learning about spatial features. Therefore, in this paper, we propose a new temporal contrast learning framework that simultaneously imparts temporal consistency, spatial consistency, and spatio-temporal consistency information to the model about the actions by performing temporal transformations and cross-contrasts on the samples. The framework of this method is shown in Figure 3.

In the spatio-temporal contrast task, a spatial augmentation module is introduced in this paper. This module consists of a series of spatial transformation operations on the 3D skeleton, including random flip, random rotation, Gaussian noise, Gaussian filtering and coordinate masking, which can effectively expand the spatial diversity of the skeleton samples. As shown in Figure 3, given a training sample X_O, its corresponding complete sample $X = S\left[r\right]$ is obtained from the sample bank according to its original sample number r. Input X_O and X into the spatial augmentation module, respectively, and obtain the corresponding spatially augmented sequences X_{O_aug} and X_{aug} for both after a series of spatial transformations.

The spatio-temporal contrast task assumes that these four samples describing the same action process have common spatio-temporal features although they have different distributions in the temporal and spatial domains. By performing cross-contrast learning among these four samples, the model can be supervised to learn feature representations with spatio-temporal consistency.

Specifically, the spatio-temporal feature extraction network consists of two parts, the shared encoder E_s and the momentum encoder E_m, which have identical structures and both perform feature extraction on samples. Among them, the shared encoder E_s receives the supervision signal generated by the back propagation of the loss function to update the weights, while the initial values of the weights in the momentum encoder E_m are the same as E_s, but instead of gradient descent during the training process, the momentum approach is adopted to update the weights according to the parameters of E_s in the following way:

$$\mathrm{param}\left(E_m, i+1\right) = m * \mathrm{param}\left(E_m, i\right) + \left(1-m\right) * \mathrm{param}\left(E_s, i+1\right), \qquad (7)$$

where $param(E,i)$ denotes the weight of model E at the i^{th} iteration and m is the momentum coefficient.

As shown in Figure 3, after the samples have passed through the spatio-temporal feature extraction network E, the resulting features are fed into the contrast mapping module CM. The contrast mapping module CM is a multi-layer perceptron that can further map the features extracted by E into the contrast learning feature space. The incomplete sequence X_O is input to E and CM in turn, and its feature vector F is output, which consists of two parts, the query vector F_{query} and the key vector F_{key}. The encoding process is given by the following equation:

$$F = CM\left(E\left(X_O\right)\right) = \left\{F_{query}, F_{key}\right\} = \left\{CM\left(E_s\left(X_O\right), CM\left(E_m\left(X_O\right)\right)\right)\right\}, \qquad (8)$$

where F_{query} is encoded by the shared encoder E_s and F_{key} is encoded by the momentum encoder E_m. Similarly, as shown in Figure 3, the complete sequence X, the partially enhanced sequence X_{O_aug}, and the complete enhanced sequence X_{aug} are input into E and CM, respectively, to obtain $F_{complete}$ consisting of $F_{complete,query}$ and $F_{complete,key}$, F_{aug} consisting of $F_{aug,query}$ and $F_{aug,key}$, and $F_{complete\_aug,query}$ and $F_{complete\_aug,key}$.

After obtaining the feature vectors of each of the four samples, the loss is generated by cross-contrast learning. The cross-prediction module CP is a multi-layer perceptron that predicts the projection of the input features in the contrast learning space. Given a feature pair $\langle F_1, F_2\rangle$, $F_{1,query}$ is input to CP and the output obtained is compared with $F_{2,key}$. At the same time, input $F_{2,query}$ to CP as well, and the obtained output is compared with $F_{1,key}$. The contrast learning loss of the feature pair $\langle F_1, F_2\rangle$ is given by the following equation:

$$\operatorname{ctr}\left(F_1, F_2\right) = -2 \cdot \left(\frac{\left\langle CP\left(F_{1,query}\right), F_{2,key}\right\rangle}{\left\|CP\left(F_{1,query}\right)\right\|_2 \left\|F_{2,key}\right\|_2} + \frac{\left\langle CP\left(F_{2,query}\right), F_{1,key}\right\rangle}{\left\|CP\left(F_{2,query}\right)\right\|_2 \left\|F_{1,key}\right\|_2}\right), \qquad (9)$$

where $\langle a, b\rangle$ denotes the inner product of two vectors a and b. Applying Equation (9) for contrast learning between F and $F_{complete}$ forces the features to be invariant in time, and similarly between F_{aug} and $F_{complete\_aug}$; contrast learning between F and F_{aug} forces the features to be invariant in space, and similarly between $F_{complete}$ and $F_{complete\_aug}$ between F_{aug} and $F_{complete\_aug}$ as well. In addition, compared to the incomplete action sequence X_O, the spatially enhanced complete sequence X_{aug} transforms in both the temporal and spatial domains, and thus contrast learning between F and $F_{complete\_aug}$ can guide the model to learn spatio-temporal invariant information, and similarly between $F_{complete}$ and F_{aug}. In summary, the loss L_C of the spatio-temporal contrast task is given by the following equation:

$$\begin{aligned} L_C = {} & \frac{1}{N}\sum_{i=1}^{N}\left(ctr\left(F_i, F_{i,complete}\right) + \mathrm{ctr}\left(F_{i,aug}, F_{i,complete\_aug}\right)\right) \\ & + \frac{1}{N}\sum_{i=1}^{N}\left(ctr\left(F_i, F_{i,aug}\right) + \mathrm{ctr}\left(F_{i,complete}, F_{i,complete\_aug}\right)\right) \\ & + \frac{1}{N}\sum_{i=1}^{N}\left(\alpha * \mathrm{ctr}\left(F_i, F_{i,complete\_aug}\right) + \beta * ctr\left(F_{i,aug}, F_{i,complete}\right)\right) \end{aligned} \tag{10}$$

where N is the batch size and α and β are two coefficients for adjusting the weight of losses.

In this task, the shared encoder E_s and the momentum encoder E_m in the spatio-temporal feature extraction network E extract the query features and key features of samples, respectively, and the spatio-temporal contrast loss L_C passes the supervision signal through back propagation, while supervising the training of the shared encoder E_s, the contrast mapping module CM and the cross-prediction module CP, and updating the weights of E_m by momentum, so as to introduce the temporal consistency, spatial consistency and spatio-temporal consistency information in 3D actions into the feature representation of the encoder.

3.4 Spatio-Temporal Feature Extraction Network Based on Graph Convolution Recurrent

The human action consists of the change of joint points in time and space, and the rich spatio-temporal information in the 3D skeleton sequence is the key to describe the action semantics. A single recurrent neural network or graph convolutional network has certain shortcomings. In this paper, we propose a spatio-temporal feature extraction network (STFEN) based on graph convolutional recurrent, which combines graph convolutional and recurrent neural networks to extract local static spatial features, inter-frame motion temporal features and global motion spatio-temporal features of 3D action sequences in a hierarchical manner while retaining the network's sensitivity to time-domain information. The structure of this network is shown in Figure 4.

The proposed spatio-temporal feature extraction network consists of two heterogeneous network modules in a hierarchy of spatial feature extraction and temporal feature transfer. The spatial feature extraction module consists of a number of graph convolutional networks with shared weights. In this module, graphs are constructed on the input set of joint point coordinates with the natural connection of the human skeleton as a priori. Specifically, the elements in the adjacency matrix A of the

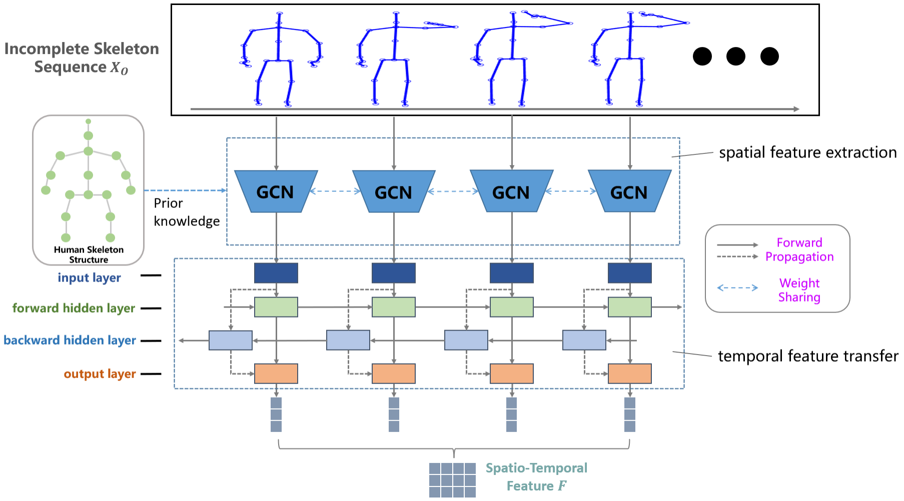


Figure 4. Schematic diagram of the spatio-temporal feature extraction network. The network consists of two modules, the spatial feature extraction and the temporal feature transfer, to extract the spatial and temporal features of the 3D skeleton sequence in a hierarchical manner.

graph are calculated as follows:

$$A_{i,j} = \begin{cases} 1, & \text{if connect}(i,j), \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where $connect(i,j)$ indicates that there is a skeletal connection between the i^{th} and j^{th} joint point in the human body structure. The set of 3D joint point coordinates of each frame is input into the graph convolution network, and the local stationary spatial features are extracted by graph convolution of the input information according to the adjacency matrix A. The local spatial feature $f_{spatial,t}$ for the t^{th} frame is calculated as follows:

$$f_{spatial,t} = GCN\left(x_t, A\right). \tag{12}$$

After obtaining all the single-frame local spatial features in the sample, these features are input into the temporal domain feature transfer module for contextual information extraction. The temporal domain feature transfer module consists of a bidirectional recurrent neural network. After obtaining the stationary spatial features of a single frame, they are input into the forward hidden layer and the backwards hidden layer for information propagation in temporal order and temporal inverse order, respectively, and the output of the states from the forward hidden layer and the backwards hidden layer is concatenated as the features of this hidden layer. After the temporal domain feature transfer, the motion features incorporating the contextual information are obtained. The t^{th} frame feature is computed as the following equation:

$$\begin{aligned} f_{motion,t} &= concate(\overrightarrow{h_t}, \overleftarrow{h_t}) \\ &= concate(\overrightarrow{W_{xh}} f_{spatial,t} + \overrightarrow{W_{hh}}\overrightarrow{h_t} + \overrightarrow{b_h}, \overleftarrow{W_{xh}} f_{spatial,t} + \overleftarrow{W_{hh}}\overleftarrow{h_t} + \overleftarrow{b_h}), \end{aligned} \tag{13}$$

where *concate* $(\cdot)$ denotes the join operation, $\overrightarrow{W_{xh}}$, $\overrightarrow{W_{hh}}$, $\overrightarrow{b_h}$ denotes the weight and bias in the forward hidden layer, and $\overleftarrow{W_{xh}}$, $\overleftarrow{W_{hh}}$, $\overleftarrow{b_h}$ denote the weights and biases in the backwards hidden layer. Finally, the single-frame motion features passed through the temporal domain features are concatenated to obtain the spatio-temporal feature F of the whole 3D action sequence. The feature contains rich spatio-temporal information and has powerful action semantic description capability.

4 EXPERIMENTS

4.1 Datasets and Experimental Settings

Following previous studies on 3D action prediction [30, 31, 32, 33, 34, 35], the proposed method is evaluated on the NTU RGB + D dataset [44]. NTU RGB + D dataset is a large-scale multimodal human action recognition dataset containing 60 action categories and 56 800 skeleton sequences. The recordings were performed by 40 volunteers and captured with the Microsoft Kinect v2 sensor. Each action is captured by 3 cameras at the same time, those have the same height but different horizontal angles: −45°, 0° and 45°. Two evaluation benchmarks are provided for this dataset:

1. Cross-Subject(CS): The dataset is divided into a training set and a testing set by subject, where the training set and the testing set each contains 20 sujects. For this evaluation, the training and testing sets have 40 320 and 16 560 samples, respectively.
2. Cross-View(CV): The dataset is divided by camera, where samples from cameras 2 and 3 are used for the training set while samples from camera 1 are used for the testing set. For this evaluation, the training and testing sets have 37 920 and 18 960 samples, respectively.

SYSU 3D HOI dataset [45] contains 12 categories of actions performed by 40 volunteers, with a total of 480 samples. All these actions are human-object interactions, captured by a Kinect camera. Since the skeleton data cannot represent the manipulated objects and some actions have the same manipulated objects and motions, it is more difficult to predict 3D actions on this dataset. We adopt the cross-subject criterion provided by the authors to evaluate the proposed method. In this setting, samples performed by half of the subjects are used as the training set and the remaining half as the testing set. The authors provided 30 random splits, we evaluate the model under each split separately, and finally report the average accuracy.

Self-Supervised Tasks		Feature Encoder	Average Prediction Accuracy	
SD	STC		Cross-Subject	Cross-View
×	×	STFEN	43.89	47.50
√	×	STFEN	45.71	46.80
×	√	STFEN	49.09	48.85
√	√	STFEN	50.19	55.13

Table 2. Self-supervised tasks ablation experiment results (%) on NTU RGB + D

The specific implementation of the proposed self-supervised learning method uses a multi-task learning framework. In addition to the state discrimination task and the temporal contrast task described in Section 3, a motion prediction task is introduced with reference to the approach in the self-supervised 3D action recognition study [10]. For the incomplete action sequence input X_O, the corresponding complete sequence X_P is generated based on its encoded features:

$$X_P = FC\left(GRU\left(F\right)\right) + X_O. \tag{14}$$

The loss calculation for the motion prediction task is then obtained as follows:

$$L_M = \frac{\sum_{i=1}^{N} \left\| X_{P,i} - X_i \right\|_2^2}{N}, \tag{15}$$

where N is the batch size.

The proposed spatio-temporal feature extraction network consists of two parts with the same structure, the shared encoder E_s and the momentum encoder E_m, where E_s is used for feature extraction in the state discrimination task, the motion prediction task, and the encoding of the query feature vector in the spatio-temporal contrast task, and E_m is used for the encoding of the key feature vector in the spatio-temporal contrast task. In the multi-task self-supervised training, E_s receives the supervision signal generated by the losses of each self-supervised task to optimize weights, and the total loss L_{self} are calculated as follows:

$$L_{self} = aL_D + bL_C + cL_O + dL_M. \tag{16}$$

Since different losses have different ranges of value distribution, a, b, c and d are introduced to balance the magnitude of losses for each self-supervised task, which are set to 0.01, 0.1, 1 and 1 in the experiments. The initial values of weights in E_m are the same as E_s, but instead of receiving the back propagation generated by losses, a momentum update is taken and calculated by applying Equation (7), where the value of momentum m is set to 0.9. Both α and β in Equation (10) are set to 0.01. After self-supervised training, the weights of E_s are fixed as the feature extractor for 3D action prediction task.

In the spatio-temporal feature extraction network, the graph convolutional network of the spatial feature extraction module uses a single-layer ST-GCN [46] with the number of channels set to 64, and the temporal feature transfer module is a two-layer bidirectional GRU network [47] with 300 neurons per layer. The ranking network RN is a two-layer perceptron with the structure *(600, 2 048)-batchnorm-relu-(2 048, 2)*. The contrast mapping module CM is a two-layer perceptron with the structure *(600, 2 048)-batchnorm-relu-(2 048, 1 024)*. The cross-prediction module CP is a three-layer perceptron with the structure *(1 024, 2 048)-batchnorm-relu-(2 048, 2 048)-batchnorm-relu-(2 048, 1 024)*. The FC in Equation (1) is a fully connected layer with the structure *(600, 1)*, GRU in Equation (14) is a two-layer GRU with 600 units in each layer. Adam optimizer is used to train the network with the batch size of 128. In the self-supervised training phase, the training epoch is set to 10, the initial learning rate is set to $5 \times e^{-4}$, and decays with a factor of 0.1 at epoch 6. In the action prediction training phase, the training epoch is set to 30, the initial learning rate is set to $1 \times e^{-2}$, and decays with a factor of 0.1 every 10 epochs.

4.2 Ablation Study

To validate the contribution of the self-supervised tasks and network structure proposed in this paper, the following ablation experiments are conducted on the NTU RGB + D dataset.

4.2.1 Effectiveness of Self-Supervised Tasks

The effectiveness of the state discrimination task and the spatio-temporal contrast task proposed in this paper is verified under two division criteria of cross-subject and cross-view in NTU RGB + D dataset. We train the network with only two losses in Equation (1) and Equation (15) as the baseline. The average prediction accuracy for different combinations of self-supervised tasks is shown in Table 2. Besides, "SD" represents State Discrimation and "STC" represents Spatio-Temporal Contrast.

As can be seen in Table 2, when the spatio-temporal feature extraction network proposed in this paper is used as the feature encoder, the average prediction accuracy of the model decreases by 1.10% and 6.28% under cross-subject and cross-view, respectivly, when the state discrimination task is removed. This indicates that the proposed state discrimination task can effectively introduce generic state information to the feature representation and enrich the action semantics in the features. Removing the spatio-temporal contrast task, the average prediction accuracy of the model decreases by 4.48% and 8.33% under the two divisions, respectively. This indicates that the proposed spatio-temporal contrast task can effectively train the model to learn temporal consistency, spatial consistency and spatio-temporal consistency knowledge, thus significantly improving the representation of features. When the state discrimination task and the spatio-temporal contrast task are removed simultaneously, the average prediction accuracy of the model decreases by 6.30% and 7.63% under the two divisions, respectively. This fully validates the usefulness of the two self-supervised tasks proposed in this paper for the model to learn feature representation. The prediction accuracy of the model at each observation rate under different combinations of self-supervised tasks is shown in Figure 5 and Figure 6.

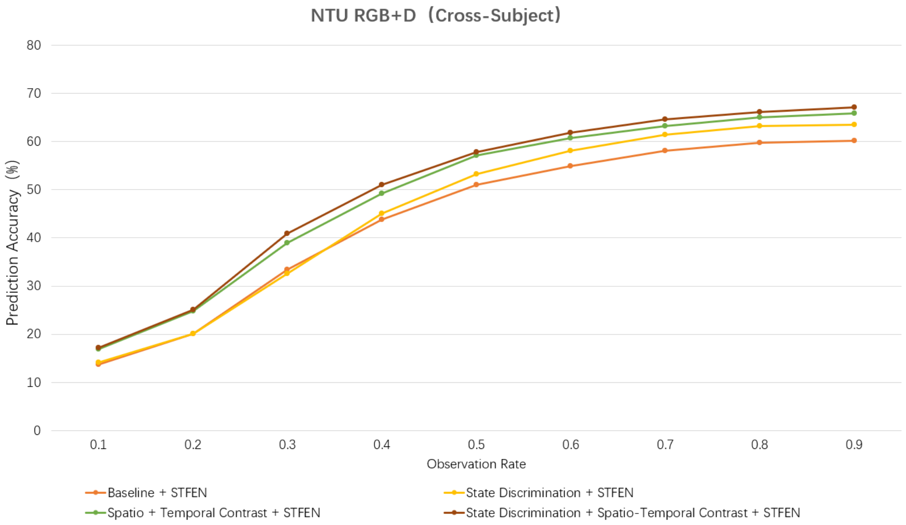


Figure 5. Comparison of prediction accuracy under different observation rates for self-supervised task combinations (Cross-Subject)

It can be visualized from Figure 5 and Figure 6 that the prediction accuracy of the model decreases under all observation rates without the state discrimination task to guide the model to learn state information across instances and categories; and the prediction accuracy of the model also decreases significantly under all observation rates without the spatio-temporal contrast task to force the encoder to learn feature representations with spatio-temporal consistency. This further demonstrates that the state discrimination task and the spatio-temporal contrast task proposed in this paper can effectively train the model to learn feature representations that contain rich action semantics.

4.2.2 Effectiveness of STFEN

We evaluate the performance on NTU RGB + D dataset with RNN and the proposed STFEN as feature encoder, respectively. The experimental results of the

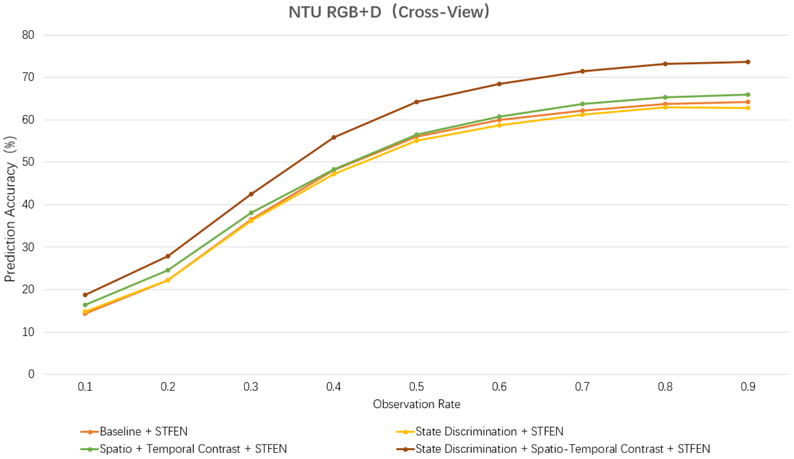


Figure 6. Comparison of prediction accuracy under different observation rates for self-supervised task combinations (Cross-View)

average prediction accuracy of the model under all observation rates are shown in Table 3.

Self-Supervised Tasks		Feature Encoder	Average Prediction Accuracy	
SD	STC		Cross-Subject	Cross-View
√	√	RNN	49.06	49.02
√	√	STFEN	50.19	55.13

Table 3. Feature encoder ablation experiment results (%) on NTU RGB + D

As shown in Table 3, after replacing the feature encoder from the spatio-temporal feature extraction network proposed in this paper with a recurrent neural network, the average prediction accuracy of the model decreases by 1.13% and 6.11% under cross-subject and cross-view, respectively. This indicates that the proposed spatio-temporal feature extraction network based on graph convolutional recurrent can capture action features containing rich spatio-temporal information and has good descriptive ability for 3D skeleton sequences. The detailed prediction accuracy of the model using different feature encoders under each observation rate is shown in Figure 7.

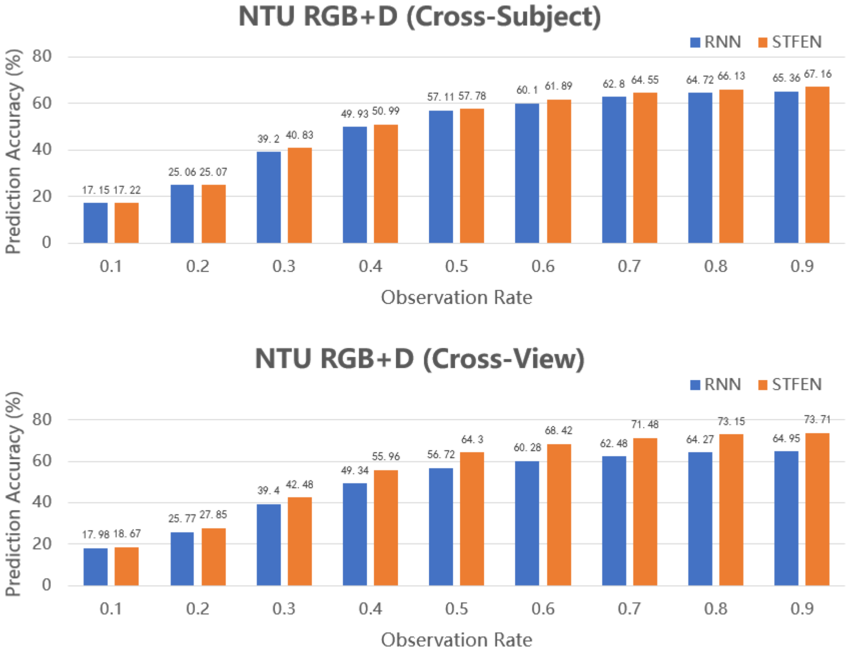


Figure 7. Comparison of prediction accuracy under different observation rates for feature encoder

Figure 7 visualizes the prediction accuracy of the model under different observation rates when using the recurrent neural network as the feature encoder and when using the spatio-temporal feature extraction network proposed in this paper as the feature encoder. It can be clearly seen that, compared to the recurrent neural network, the prediction accuracy of the model under all observation rates is somewhat improved when the spatio-temporal feature extraction network is used as the feature encoder on both divisions of the NTU RGB + D dataset. This fully illustrates the effectiveness of the spatio-temporal feature extraction network based on graph convolutional recurrent for feature encoding proposed in this paper.

4.3 Comparison with Other Methods

To the best of our knowledge, there are no other publically available studies on self-supervised 3D action prediction. Therefore, the performance of supervised 3D action prediction methods is listed in the comparison as a reference. The experimental results on NTU RGB + D dataset under two divisions are detailed in Table 4 and Table 5.

It can be seen that the prediction accuracy of the proposed self-supervised method has outperformed that of some supervised learning methods [27, 28, 48]

Methods	Backbone	Observation Ratios			
		0.2	0.4	0.6	0.8
Supervised Methods					
F-RNN-EL (ICRA 16) [27]	RNN	7.07	18.98	44.55	63.84
MTLN (CVPR 17) [48]	CNN	8.34	26.97	56.78	75.13
Local + LGN (TIP 19) [30]	CNN	32.12	63.82	77.02	82.45
CEL (TCSVT 20) [31]	RNN	35.56	54.63	67.08	72.91
HARD-Net (ECCV 20) [32]	GCN	42.39	72.24	82.99	86.75
Local + AGCN-AL(TCDS 21) [33]	GCN	38.18	71.19	82.25	86.33
GCN-EAM (ICPR 22) [34]	GCN	41.50	72.23	82.07	85.61
$S^1 + S^2$ (SPL 22) [35]	GCN	42.32	73.62	84.1	87.83
Self-Supervised Method					
Ours	RNN	25.07	50.99	61.89	66.13

Table 4. Action prediction accuracy (%) on NTU RGB + D (Cross-Subject)

Methods	Backbone	Observation Ratios			
		0.2	0.4	0.6	0.8
Supervised Methods					
CEL (TCSVT 20) [31]	RNN	37.22	57.18	69.92	75.41
HARD-Net (ECCV 20) [32]	GCN	53.15	82.87	91.34	93.71
Self-Supervised Method					
Ours	RNN	27.85	55.96	68.42	73.15

Table 5. Action prediction accuracy (%) on NTU RGB + D (Cross-View)

at each observation rate. Under two divisions of the NTU RGB + D dataset, the performance of our self-supervised method is already close to that of the supervised 3D action prediction method [31] in 2020. Compared with the latest supervised 3D action prediction methods [34, 35], the prediction accuracy of the proposed self-supervised method is about 20% lower on average at each observation rate. This fully illustrates the effectiveness of the self-supervised learning method for 3D action prediction proposed in this paper.

Methods	Backbone	Observation Ratios			
		0.2	0.4	0.6	0.8
Supervised Methods					
F-RNN-EL (ICRA 16) [27]	RNN	31.61	53.37	68.71	73.96
MTLN (CVPR 17) [48]	CNN	26.76	52.86	72.32	79.40
Local + LGN (TIP 19) [30]	CNN	58.81	74.21	82.18	84.42
Local + AGCN-AL (TCDS 21) [33]	GCN	63.46	80.93	87.92	90.38
Self-Supervised Method					
Ours	RNN	25.98	51.84	62.73	67.67

Table 6. Action prediction accuracy (%) on SYSU 3D HOI (Cross-Subject)

The experimental results on SYSU 3D HOI dataset are detailed in Table 6. It is evident that the performance gap between our self-supervised method and the supervised method is maintained within an acceptable range.

In addition, we have conducted experiments on 3D action recognition task. The experimental results on NTU RGB + D dataset under two divisions are detailed in Table 7. It can be seen that the performance of our method is competitive compared to the recent self-supervised 3D action recognition method [20].

Methods	Backbone	Cross-Subject	Cross-View
LongTGAN (AAAI 18) [9]	RNN	39.1	48.1
MS^{2}L (ACM MM 20) [10]	GCN	52.6	–
CSSL-SAR (NeurIPS Workshop 20) [12]	ResNet	52.3	62.1
PCRP (TMM 21) [13]	GRU	53.9	63.5
CAE+ (Information Sciences 21) [14]	LSTM	58.5	64.8
P&C FW-AEC (CVPR 20) [15]	GRU	50.7	76.1
CRRL (TIP 21) [18]	RNN	67.6	73.8
SkeletonCLR (CVPR 21) [16]	GCN	68.3	76.4
AimCLR (AAAI 22) [20]	GCN	64.3	**79.7**
MG-AL (TCSVT 22) [25]	GCN	64.7	68.0
Ours	RNN	**74.5**	78.9

Table 7. Comparison of self-supervised action recognition methods on NTU RGB + D

4.4 Qualitative Analysis

The confusion matrix of the proposed method on the NTU RGB + D dataset is shown in Figure 8. As can be seen from the figure, the proposed method is less able to discriminate between actions with smaller amplitude and finer granularity of motion areas, such as "drink water" and "eat meal/snack", "make a phone call" and "playing with phone". This may be due to the fact that the discriminative of these actions is the object with which the person interacts, but no object information is reflected in the skeleton sequence, and the distinction in action performance is small, resulting in less discriminative features extracted by the network.

5 CONCLUSION

In this paper, we propose a self-supervised learning method for 3D action prediction based on state discrimination and spatio-temporal contrastive graph convolution recurrent. It guides the model for cross-instance and cross-category state information learning through relative action completeness perception, and endowing the feature representation with temporal consistency, spatial consistency and spatio-temporal consistency through spatio-temporal cross-contrast learning. For the rich spatio-temporal information in 3D action sequences, a spatio-temporal feature extraction network based on graph convolution recurrent is proposed, including two modules of

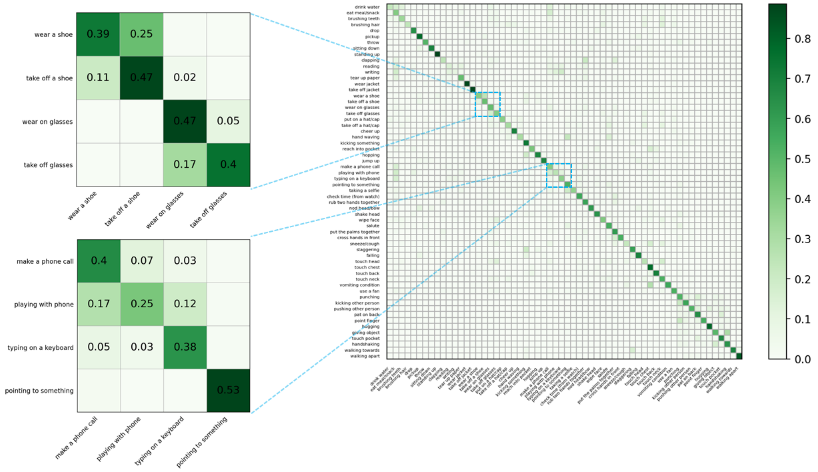


Figure 8. Comparison of prediction accuracy under different observation rates for feature encoder

spatial feature extraction and temporal feature transfer, to enrich the action semantics in the feature representation by combining the characteristics of heterogeneous networks. The proposed self-supervised learning method and the spatio-temporal feature extraction network are evaluated on two 3D action datasets, and the experimental results fully demonstrate the effectiveness of each part of the proposed method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62571464, the Shenzhen Science and Technology Projects under Grant No. JCYJ20200109143035495, and the Natural Science Foundation of Fujian Province under Grants No. 2022J011275 and 2023J01003.

REFERENCES

[1] Hinton, G. E.—Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks. Science, Vol. 313, 2006, No. 5786, pp. 504–507, doi: 10.1126/science.1127647.

[2] Hinton, G. E.—Osindero, S.—Teh, Y. W.: A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, Vol. 18, 2006, No. 7, pp. 1527–1554, doi: 10.1162/neco.2006.18.7.1527.

[3] Wu, X.—Zhao, J.—Wang, R.: Anticipating Future Relations via Graph Growing for Action Prediction. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, No. 4, pp. 2952–2960, doi: 10.1609/aaai.v35i4.16402.

[4] Liu, C.—Gao, Y.—Li, Z.—Du, C.—Liu, F.—Shi, X.: Action Prediction Network with Auxiliary Observation Ratio Regression. 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428266.

[5] Huang, J.—Li, N.—Li, T.—Liu, S.—Li, G.: Spatial-Temporal Context-Aware Online Action Detection and Prediction. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 30, 2020, No. 8, pp. 2650–2662, doi: 10.1109/TCSVT.2019.2923712.

[6] Tao, L.—Wang, X.—Yamasaki, T.: An Improved Inter-Intra Contrastive Learning Framework on Self-Supervised Video Representation. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, 2022, No. 8, pp. 5266–5280, doi: 10.1109/TCSVT.2022.3141051.

[7] Zhu, Y.—Shuai, H.—Liu, G.—Liu, Q.: Self-Supervised Video Representation Learning Using Improved Instance-Wise Contrastive Learning and Deep Clustering. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, 2022, No. 10, pp. 6741–6752, doi: 10.1109/TCSVT.2022.3169469.

[8] Han, J.—Shao, L.—Xu, D.—Shotton, J.: Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. IEEE Transactions on Cybernetics, Vol. 43, 2013, No. 5, pp. 1318–1334, doi: 10.1109/TCYB.2013.2265378.

[9] Zheng, N.—Wen, J.—Liu, R.—Long, L.—Dai, J.—Gong, Z.: Unsupervised Representation Learning with Long-Term Dynamics for Skeleton Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, No. 1, pp. 2644–2651, doi: 10.1609/aaai.v32i1.11853.

[10] Lin, L.—Song, S.—Yang, W.—Liu, J.: MS2L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. Proceedings of the 28th ACM International Conference on Multimedia (MM '20), 2020, pp. 2490–2498, doi: 10.1145/3394171.3413548.

[11] Si, C.—Nie, X.—Wang, W.—Wang, L.—Tan, T.—Feng, J.: Adversarial Self-Supervised Learning for Semi-Supervised 3D Action Recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12352, 2020, pp. 35–51, doi: 10.1007/978-3-030-58571-6_3.

[12] Gao, X.—Yang, Y.—Du, S.: Contrastive Self-Supervised Learning for Skeleton Action Recognition. In: Bertinetto, L., Henriques, J. F., Albanie, S., Paganini, M., Varol, G. (Eds.): NeurIPS 2020 Workshop on Pre-Registration in Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 148, 2021, pp. 51–61, `http://proceedings.mlr.press/v148/gao21a/gao21a.pdf`.

[13] Xu, S.—Rao, H.—Hu, X.—Cheng, J.—Hu, B.: Prototypical Contrast and Reverse Prediction: Unsupervised Skeleton Based Action Recognition. IEEE Transactions on Multimedia, Vol. 25, 2021, pp. 624–634, doi: 10.1109/TMM.2021.3129616.

[14] Rao, H.—Xu, S.—Hu, X.—Cheng, J.—Hu, B.: Augmented Skeleton Based Con-

trastive Action Learning with Momentum LSTM for Unsupervised Action Recognition. Information Sciences, Vol. 569, 2021, pp. 90–109, doi: 10.1016/j.ins.2021.04.023.

[15] Su, K.—Liu, X.—Shlizerman, E.: PREDICT & CLUSTER: Unsupervised Skeleton Based Action Recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9628–9637, doi: 10.1109/CVPR42600.2020.00965.

[16] Li, L.—Wang, M.—Ni, B.—Wang, H.—Yang, J.—Zhang, W.: 3D Human Action Representation Learning via Cross-View Consistency Pursuit. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4739–4748, doi: 10.1109/CVPR46437.2021.00471.

[17] Su, Y.—Lin, G.—Wu, Q.: Self-Supervised 3D Skeleton Action Representation Learning with Motion Consistency and Continuity. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13308–13318, doi: 10.1109/ICCV48922.2021.01308.

[18] Wang, P.—Wen, J.—Si, C.—Qian, Y.—Wang, L.: Contrast-Reconstruction Representation Learning for Self-Supervised Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, Vol. 31, 2022, pp. 6224–6238, doi: 10.1109/TIP.2022.3207577.

[19] Thoker, F. M.—Doughty, H.—Snoek, C. G. M.: Skeleton-Contrastive 3D Action Representation Learning. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 1655–1663, doi: 10.1145/3474085.3475307.

[20] Guo, T.—Liu, H.—Chen, Z.—Liu, M.—Wang, T.—Ding, R.: Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-Supervised Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, No. 1, pp. 762–770, doi: 10.1609/aaai.v36i1.19957.

[21] Chen, Z.—Liu, H.—Guo, T.—Chen, Z.—Song, P.—Tang, H.: Contrastive Learning from Spatio-Temporal Mixed Skeleton Sequences for Self-Supervised Skeleton-Based Action Recognition. CoRR, 2022, doi: 10.48550/arXiv.2207.03065.

[22] Wu, B.—Wu, M.—Ji, H.—Shen, L.: Which One Is Better? Self-Supervised Temporal Coherence Learning for Skeleton Based Action Recognition. 2022 IEEE International Joint Conference on Biometrics (IJCB), 2022, pp. 1–9, doi: 10.1109/IJCB54206.2022.10007979.

[23] Pang, C.—Lu, X.—Lyu, L.: Skeleton-Based Action Recognition Through Contrasting Two-Stream Spatial-Temporal Networks. IEEE Transactions on Multimedia, Vol. 25, 2023, pp. 8699–8711, doi: 10.1109/TMM.2023.3239751.

[24] Zhao, Z.—Chen, G.—Lin, Y.: Temporal-Masked Skeleton-Based Action Recognition with Supervised Contrastive Learning. Signal, Image and Video Processing, Vol. 17, 2023, No. 5, pp. 2267–2275, doi: 10.1007/s11760-022-02442-6.

[25] Yang, Y.—Liu, G.—Gao, X.: Motion Guided Attention Learning for Self-Supervised 3D Human Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, 2022, No. 12, pp. 8623–8634, doi: 10.1109/TCSVT.2022.3194350.

[26] Hu, J. F.—Zheng, W. S.—Ma, L.—Wang, G.—Lai, J.: Real-Time RGB-D Activity Prediction by Soft Regression. In: Leibe, B., Matas, J., Sebe, N., Welling, M.

(Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9905, 2016, pp. 280–296, doi: 10.1007/978-3-319-46448-0_17.

[27] Jain, A.—Singh, A.—Koppula, H. S.—Soh, S.—Saxena, A.: Recurrent Neural Networks for Driver Activity Anticipation via Sensory-Fusion Architecture. 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 3118–3125, doi: 10.1109/ICRA.2016.7487478.

[28] Ke, Q.—Liu, J.—Bennamoun, M.—Rahmani, H.—An, S.—Sohel, F.—Boussaid, F.: Global Regularizer and Temporal-Aware Cross-Entropy for Skeleton-Based Early Action Recognition. In: Jawahar, C. V., Li, H., Mori, G., Schindler, K. (Eds.): Computer Vision – ACCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11364, 2019, pp. 729–745, doi: 10.1007/978-3-030-20870-7_45.

[29] Liu, J.—Shahroudy, A.—Wang, G.—Duan, L. Y.—Kot, A. C.: Skeleton-Based Online Action Prediction Using Scale Selection Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, 2020, No. 6, pp. 1453–1467, doi: 10.1109/TPAMI.2019.2898954.

[30] 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, Vol. 29, 2019, pp. 959–970, doi: 10.1109/TIP.2019.2937757.

[31] Weng, J.—Jiang, X.—Zheng, W. L.—Yuan, J.: Early Action Recognition with Category Exclusion Using Policy-Based Reinforcement Learning. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 30, 2020, No. 12, pp. 4626–4638, doi: 10.1109/TCSVT.2020.2976789.

[32] Li, T.—Liu, J.—Zhang, W.—Duan, L.: HARD-Net: Hardness-AwaRe Discrimination Network for 3D Early Activity Prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12356, 2020, pp. 420–436, doi: 10.1007/978-3-030-58621-8_25.

[33] Li, G.—Li, N.—Chang, F.—Liu, C.: Adaptive Graph Convolutional Network with Adversarial Learning for Skeleton-Based Action Prediction. IEEE Transactions on Cognitive and Developmental Systems, Vol. 14, 2022, No. 3, pp. 1258–1269.

[34] Liu, C.—Zhao, X.—Yan, Z.—Jiang, Y.—Shi, X.: A Graph Convolutional Network with Early Attention Module for Skeleton-Based Action Prediction. 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 1266–1272, doi: 10.1109/ICPR56361.2022.9956108.

[35] Liu, C.—Zhao, X.—Li, Z.—Yan, Z.—Du, C.: A Novel Two-Stage Knowledge Distillation Framework for Skeleton-Based Action Prediction. IEEE Signal Processing Letters, Vol. 29, 2022, pp. 1918–1922, doi: 10.1109/LSP.2022.3204190.

[36] Gepshtein, S.—Wang, Y.—He, F.—Diep, D.—Albright, T. D.: A Perceptual Scaling Approach to Eyewitness Identification. Nature Communications, Vol. 11, 2020, No. 1, Art. No. 3380, doi: 10.1038/s41467-020-17194-5.

[37] Li, Y.—Hu, P.—Liu, Z.—Peng, D.—Zhou, J. T.—Peng, X.: Contrastive Clustering. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, No. 10, pp. 8547–8555, doi: 10.1609/aaai.v35i10.17037.

[38] Li, Y.—Yang, M.—Peng, D.—Li, T.—Huang, J.—Peng, X.: Twin Contrastive Learning for Online Clustering. International Journal of Computer Vision, Vol. 130, 2022, No. 9, pp. 2205–2221, doi: 10.1007/s11263-022-01639-z.

[39] Xu, B.—Shu, X.—Song, Y.: X-Invariant Contrastive Augmentation and Representation Learning for Semi-Supervised Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, Vol. 31, 2022, pp. 3852–3867, doi: 10.1109/TIP.2022.3175605.

[40] Xu, B.—Shu, X.—Zhang, J.—Dai, G.—Song, Y.: Spatiotemporal Decouple-and-Squeeze Contrastive Learning for Semisupervised Skeleton-Based Action Recognition. IEEE Transactions on Neural Networks and Learning Systems, Vol. 35, 2024, No. 8, doi: 10.1109/TNNLS.2023.3247103.

[41] Shu, X.—Xu, B.—Zhang, L.—Tang, J.: Multi-Granularity Anchor-Contrastive Representation Learning for Semi-Supervised Skeleton-Based Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 6, pp. 7559–7576, doi: 10.1109/TPAMI.2022.3222871.

[42] Xu, B.—Shu, X.: Pyramid Self-Attention Polymerization Learning for Semi-Supervised Skeleton-Based Action Recognition. CoRR, 2023, doi: 10.48550/arXiv.2302.02327.

[43] Yan, R.—Xie, L.—Shu, X.—Zhang, L.—Tang, J.: Progressive Instance-Aware Feature Learning for Compositional Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 8, pp. 10317–10330, doi: 10.1109/TPAMI.2023.3261659.

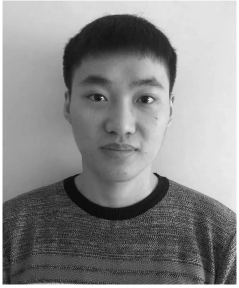
[44] Shahroudy, A.—Liu, J.—Ng, T. T.—Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019, doi: 10.1109/CVPR.2016.115.

[45] Hu, J. F.—Zheng, W. S.—Lai, J.—Zhang, J.: Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2017, No. 11, pp. 2186–2200, doi: 10.1109/TPAMI.2016.2640292.

[46] Yan, S.—Xiong, Y.—Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, No. 1, pp. 7444–7452, doi: 10.1609/aaai.v32i1.12328.

[47] Schuster, M.—Paliwal, K. K.: Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing, Vol. 45, 1997, No. 11, pp. 2673–2681, doi: 10.1109/78.650093.

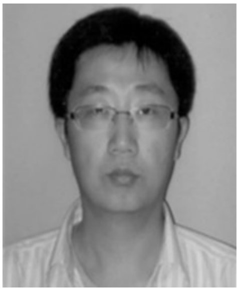
[48] Ke, Q.—Bennamoun, M.—An, S.—Sohel, F.—Boussaid, F.: A New Representation of Skeleton Sequences for 3D Action Recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4570–4579, doi: 10.1109/CVPR.2017.486.



Peng Liu is currently pursuing a graduation degree with the Department of Computer Science and Technology, Xiamen University, Fujian, China. His research interests include computer vision and deep learning.

Yifan Wang is currently pursuing a graduation degree with the Department of Computer Science and Technology, the Xiamen University, Fujian, China. His research interests include computer vision and machine learning.

Qicong Wang received his Ph.D. degree in information and communication engineering from the Zhejiang University, Hangzhou, China. He is currently Associate Professor at the Department of Computer Science and Technology, the Xiamen University, Xiamen, China. His research interests include computer vision, machine learning, big data analytic.



Chong Zhao received his Ph.D. degree in the Department of Computer Science and Engineering from the Chinese University of Hong Kong. He is currently Assistant Professor in the Department of Computer Science, the Xiamen University. His research interests include geometry processing, computer graphics and computer vision.

Yan Chen received her M.Sc. degree from the Brunel University London, U.K., in 2018. She is currently a Lecturer at the College of Business and Management, the Xiamen Huaxia University, Xiamen, China. Her current research interest includes data analysis and modeling.

Man Qi is a Reader in Computing at Canterbury Christ Church University. Her research interests are in cybersecurity, data intelligence, IoT and human-computer interaction (HCI). She has published over 80 research papers, including more than 30 journal, and serves on the editorial board of five international journals. She has been the Ph.D. external examiner for many Universities in Australia and the U.K. She has served as chair/program committee member for around 50 international conferences and been a long-term reviewer for many international journals. She is a Fellow of the British Computer Society (FBCS) and Fellow of the Higher Education Academy (FHEA).