Spatial Temporal Enhanced Contrastive and Pretext Learning for Skeleton-based Action Representation

Yiwen Zhan* ZHANYIWEN2016@BUPT.EDU.CN
Yuchen Chen* CYC99@BUPT.EDU.CN
Pengfei Ren RPF@BUPT.EDU.CN
Haifeng Sun† HFSUN@BUPT.EDU.CN
Jingyu Wang WANGJINGYU@BUPT.EDU.CN
QiQi QIQI8266@BUPT.EDU.CN
Jianxin Liao JXLBUPT@GMAIL.COM
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

In this paper, we focus on unsupervised representation learning for skeleton-based action recognition. The critical issue of this task is extracting discriminative spatial-temporal information from skeleton sequences to form action representation. To better solve this, we propose a novel unsupervised framework named contrastive-pretext spatial-temporal network (CP-STN), aiming to achieve accurate action recognition by better exploiting discriminative spatial-temporal enhanced features from massive unlabeled data. We combine contrastive and pretext tasks learning paradigms in one framework by using asymmetric spatial and temporal augmentations to enable network extracting discriminative representations with spatial-temporal information fully. Furthermore, graph-based convolution is used as the backbone to explore natural spatial-temporal graph information in skeleton data. Extensive experimental results show that our CP-STN significantly boosts the performance of existing skeleton-based action representations learning networks and achieves state-of-the-art accuracy on two challenging benchmarks in both unsupervised and semi-supervised settings.

Keywords: Action Recognition, Contrastive Learning, Spatial-temporal Feature Extraction

1. Introduction

As an essential problem in computer vision, human action recognition(Carreira and Zisserman (2017); Song et al. (2016); Liu et al. (2019); Chen et al. (2020)), plays a crucial role in video understanding, video surveillance and human-computer interaction. Due to the compact and effective skeletal representation of human body and its background-invariant feature, skeleton-based action recognition draws broad attention (Du et al. (2015a); Caetano et al. (2019)).

. *Equal contribution
. †Corresponding author

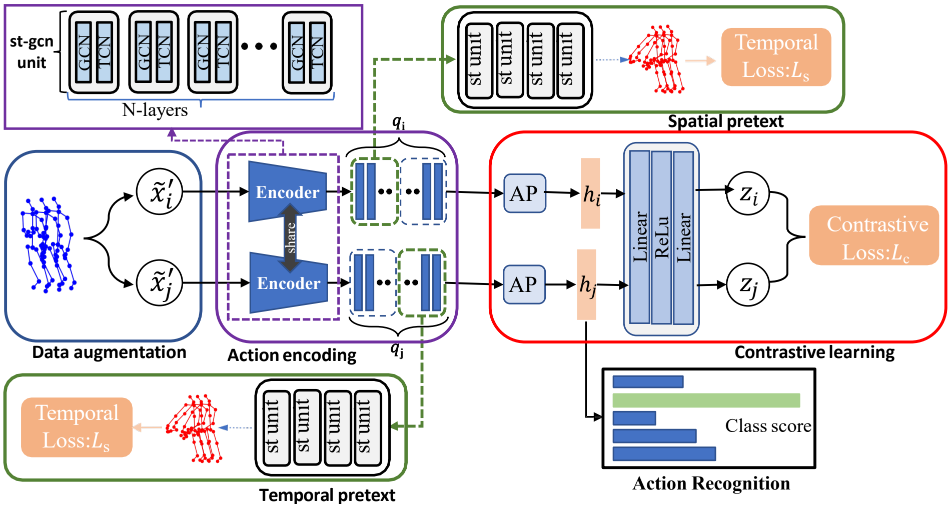


Figure 1: The overall framework of contrastive-pretext spatial-temporal Network (CP-STN). The CP-STN consists of four parts: (1) Data augmentations: sample and does spatial and temporal augmentations separately to the original skeleton sequences. (2) Action encoding: use shared spatial-temporal graph convolution units to extract representations (3) Contrastive learning: learn consistent features of spatial and temporal augmented features learned by the encoder. AP refers to time average pooling. (4) Spatial and temporal pretexts: assist contrastive learning in learning fine-grained spatial-temporal information

While recent methods have shown remarkable success in this area, they rely on strong supervision with large number of annotated training data. However, the process of annotating training data can be tedious and expensive, which might face uncertain labeling or mislabelling challenges due to the high inter-class similarity of actions. Thus, exploiting unlabeled data to learn action representations attracts increasing attention. To achieve better representation learning without labeling, a method should learn discriminative classification representations from data. As for skeleton-based action recognition, spatial-temporal information which is occupied in the sequences, can benefit the recognition performances (Liu et al. (2020a); Li et al. (2019); Shi et al. (2019); Si et al. (2019)). Recently, there are a few attempts on skeleton-based action representation learning. Most of them learn action representations via contrastive learning (Rao et al. (2020); Lin et al. (2020)) or sequential reconstruction (Su et al. (2020)), and achieve good performance. For contrastive learning methods, they try to pull closer two augmented samples of one sequence regardless of others, trying to learn discriminative classifications features of the inputs for classification. It can perform well on most of actions. However, confusing actions such as actions putting on and taking off a hat, which are opposite in the time, may be misjudged, for neglecting fine-grained spatial-temporal information. Another type of method to achieve self-supervision is sequential reconstructions. Precisely, this method(Su et al. (2020)) reconstructs the input from the learned feature that forces the predicted sequences to get closer to the original one, which can encourage the representation to encode more detailed information. However, it may overlook the most important spatial-temporal information and lead to noisy representations. So it is essential to build a network that learns discriminative classification information, which is contained with fine-grained spatial-temporal information. By combining the advantages of sequential reconstruction and augmented sequence contrast methods, we propose a novel action representation learning method called contrastive-pretext spatial-temporal network (CP-STN).

The proposed CP-STN architecture is shown in Figure 1. It follows contrastive learning framework, trying to extract features in an unsupervised style by maximizing consistency between different augmented data features via a contrastive loss. We combine spatial-temporal pretext tasks in the framework. Specifically, asymmetric spatial and temporal augmentations are applied to data, then two features generated by encoder will feed into decoder and reconstruct the input.

To better cooperate with our proposed paradigm, we need a more powerful backbone capable of modeling spatial-temporal features. We represent a skeleton as a graph structure to characterize the spatial relations between joints in each frame based on graph convolution networks (GCN)(Yan et al. (2018)). Temporal convolution network (TCN)(Kim and Reiter (2017)) is used to capture temporal information. GCN and TCN are combined into an st-gcn unit that can learn spatial and temporal information synchronously. The encoder we used is composed of several st-gcn units.

To verify the proposed CP-STN, extensive experiments are performed on three large-scale datasets: NTU RGB+D 60(Shahroudy et al. (2016)), NTU RGB+D 120(Liu et al. (2019)) and NW-UCLA(Wang et al. (2014)). Our model achieves state-of-the-art performance on these datasets. The contributions of our work are listed as follows:

- We propose an unsupervised action representation learning paradigm named CP-STN for skeleton-based action recognition. The proposed CP-STN combine pretext tasks and contrastive learning to extract discriminative classification information containing fine-grained spatial temporal information.

- We use graph-based architecture to extract features of skeleton data, which draws the limitations of previous unsupervised methods that ignore the bones' connection, and adding temporal connection for each frame. To our knowledge, it is the first work to use spatial-temporal graph structure to model skeleton data in unsupervised setting.

- Experiments on NTU RGB+D 60, NTU RGB+D 120 and NW-UCLA datasets show that our proposed method outperforms other unsupervised state-of-art methods by a large margin.

2. Related Work

2.1. Skeleton-based Action Representation Learning

For skeleton-based action recognition, unsupervised representation learning is an emerging area. Zheng et al. (2018) proposes a generative adversarial network encoder-decoder model to learn the action representation for downstream action classification. Su et al. (2020) proposes an LSTM-based encoder and a weaken decoder to regenerate noised data in a self-supervision manner and use KNN to achieve the purpose of classification. Rao et al. (2020)

first combines the task with contrastive learning method to extract the global representation by using different augmentations of input data. Lin et al. (2020) integrates multiple self-supervised tasks including contrastive learning(Chen et al. (2020)), motion prediction and jigsaw puzzle recognition to learn features. However, these methods pay less attention to the co-occurrence relationship between spatial and temporal information.

2.2. Spatial-temporal Information Extraction

The existing methods explore different models to learn spatial and temporal features. A spatial-temporal attention model based on LSTM is used by Song et al. (2016) to select spatial and temporal features. Ke et al. (2017) employs the Convolutional Neural Networks (CNNs) to learn spatial-temporal features from skeletons. Yan et al. (2018) first applies graph convolutional networks (GCN) for action recognition. They construct a spatial graph based on joints' natural connections in the human body and add the temporal edges between corresponding joints in consecutive frames. To further increase the model's flexibility for graph construction and bring more generality to adapt to various data samples. Shi et al. (2019) proposes a two-stream network to simultaneously model both the first and the second order information and use a data-driven graph to represent inner structure of skeleton sequences. However, all these methods focus on supervised learning. Our method combine spatial-temporal information extraction with unsupervised setting.

3. Methods

In this section, our CP-STN is introduced. We focus on skeleton-based unsupervised representation learning. This task aims to learn a powerful encoder $f(\cdot)$ from body joints without labels. Body joint is formulated as $X^m = (x_1^m, ..., x_T^m)$, containing T consecutive skeleton frames, where $x_i^m \in \mathbb{R}^{M\times J\times 3}$ is 3D coordinates of J different joints for M persons. The learned encoder is used to get the feature representations for action recognition task.

3.1. Overall Framework

Figure 1 depicts our proposal framework. First feeding two augmentations into a shared encoder which is composed of six spatial-temporal graph convolutions units (st-gcn units) to get features, and obtaining $q_i = f(\tilde{X}_i^m)$ and $q_j = f(\tilde{X}_j^m)$ where $q_i, q_j \in \mathbb{R}^{M\times C\times T\times V}$. Then, taking different half parts of learned features q_i, q_j into spatial-temporal graph neural network separately to learn fine-grained spatial and temporal information. After this, the network can accentuate the features that are adapted to spatial information or temporal information separately, while shunning irrelevant ones. Then, an time average pooling layer is used to get representations $h_i/h_j \in \mathbb{R}^d$ used for action recognition in the testing stage. Besides, it is further converted to contrastive space for contrastive learning. In the following sections,each part of our framework will be introduced in detail.

3.2. Contrastive-Pretext Learning

To better exploit both global and fine-grained spatial-temporal information. Our methods combining contrastive and pretext task learning.

3.2.1. Contrastive Learning

For contrastive learning methods, after extracting features of two types of augmented sequences from the backbone, a time average pooling is used to aggregate global action encoding representations across time. Then, there is a two-linear-layer multi-layer perceptron (MLP) used to project the representations to contrastive space. Our network learns representations by maximizing cosine similarity between the features in the contrastive space. The aim of contrastive learning is to learn global representations of the skeleton data.

3.2.2. Spatial Pretext Task

After getting feature $q_i \in \mathbb{R}^{M \times C \times T \times V}$, feeding first half of channel of q_i in the spatial pretext network, which aim to regenerate input data from the spatial augmented sequences. The augmentations for this branch consist of rotation, shear transformation and random joints perturbation which is shown in Figure 2(a)subfigure, and with definitions as below:

- Rotation: For all joints in a skeleton data, to gather the angle-invariant spatial information, we randomly choose a rotation angle from $[0, \pi/6]$ for a randomly selected axis.
- Shear: For shearing, it aims to obtain a sequence by scaling the directed distance from each point of the input to a specific line parallel to the direction according to a certain ratio in a certain direction. And the shear ratio for each directions is random chosen from [-1,1].
- Joints jittering: In order to learn position-invariant information of data, we apply Gaussian noise randomly over joints coordinate with a probability of 0.2.

3.2.3. Temporal Pretext Task

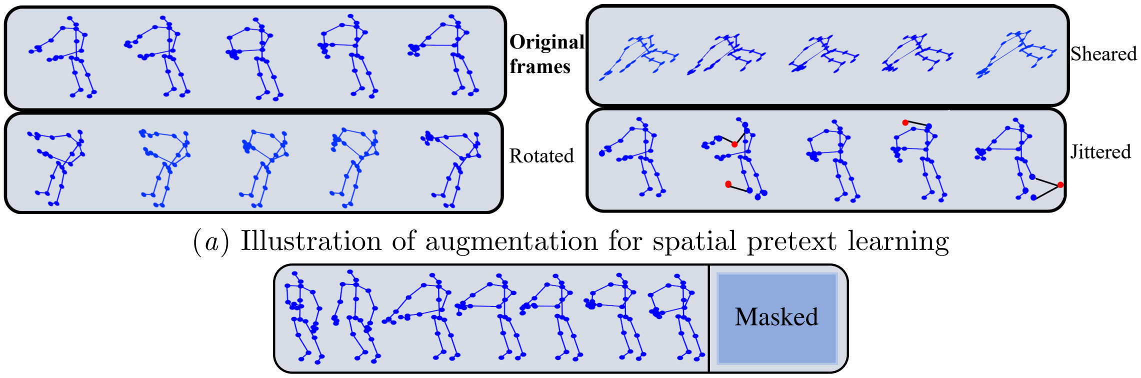
Given the past skeleton sequence, the encoder f($\cdot$) reads in parts of the input sequences $\tilde{X}_j^m$ which is illustrated in Figure 2(b)subfigure and extracts representations from inputs. The temporal pretext task receives the learned representations and reconstructs the masked part from these, getting the reconstructed sequence $\hat{X}_j^m$.

3.3. Spatial-Temporal Information Extraction Encoder

Adaptive spatial temporal convolution is always used to exploit spatio-temporal cooperation in encoder better. It aims to adaptively learn the topology of the graph for different GCN layers and skeleton samples. It can be formulated as:

$$f_{out} = \sum_{k}^{K_v} W_k f_{in}(A_k + B_k + C_k) \tag{1}$$

where A_k is the physical structure of the human body, B_k is used to learn nonadjacent connections without constraints, depending on the training data completely. And C_k is a data-driven graph to learn a unique graph for each sample by attention. Though a learnable graph structure is helpful for spatial information extraction, it relies heavily on labeled data. For B_k, it is unconstrained, which may be confused by the similar action. C_k



(a) Illustration of augmentation for spatial pretext learning

(b) Illustration of augmentation for temporal pretext learning

Figure 2: Illustration of spatial-temporal augmentations for contrastive and pretext learning

use attention to learn data-dependence relations between different joints may cause different actions with similar graph representations(Liu et al. (2020b)), which contrasts with the aim of contrastive learning. So, we argue that just using A will enough. In Section 4.3.1, we verify our assumption. After gathering structure by GCN, we use TCN to extract temporal feature, which is represented as $t(\cdot)$. We combine GCN and TCN as a st-gcn unit, which can be formulated as:

$$f_{out} = t(\sum_{k}^{K_v} W_k f_{in}(A_k)) \tag{2}$$

3.4. Training Objective

This method combines contrastive learning and spatial-temporal pretext tasks into one framework. With the representations extracted by the encoder, a classifier on the top of the encoder is used to achieve action recognition. The pretrained model is obtained in an end-to-end training strategy, the loss which composes of contrastive learning and two pretext tasks, can summarize as below:

$$L = \alpha L_c + \beta L_s + \gamma L_t \tag{3}$$

where α, β, and γ are weighting factors, and all of them are equal to one here.

3.4.1. Contrastive loss

This loss ensures representations of positive pairs $\tilde{X}_i^m$ and $\tilde{X}_j^m$ closer in the contrastive space, and identifies negative examples $\{\tilde{X}^k\}_{k\neq m}$. let z_s, z_t be the features extracted from the projection, $sim(u,v) = u^T v/||u||_2||v||_2$ defines the similarity between u and v, we define the loss for a positive pair (s,t) as below :

$$l_{s,t} = -log\frac{exp(sim(z_s,z_t)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq s]} exp(sim(z_s,z_k)/\tau)} \tag{4}$$

where $\mathit{1}_{[k \neq s]} \in \{0,1\}$ equals to 1 if and only if $k \neq s$, and the loss is computed for all the positive pairs, not only for (s,t) but also for (t,s). For all the batch N, the contrastive loss is defined as:

$$L_c = \frac{1}{N} \sum_{k=1}^{N} (l_{2k-1,2k} + l_{2k,2k+1}) \tag{5}$$

3.4.2. Spatial Loss

The spatial loss aims to accompany the contrastive learning network to learn more spatial information. The MSE loss is used to judge the regeneration accuracy to learn more detail spatial-invariant representations.

$$L_s = \sum_{i=1}^{N} \sum_{t=1}^{T} ||\hat{x}_t^i - x_t^i||_2^2 \tag{6}$$

where x_t^i is the original data, $\hat{x}_t^i$ is the regenerated data and N is the batch size.

3.4.3. Temporal Loss

For the input sequence $X^m = \{x_1^m, ..., x_T^m\}$, the masked data is $\tilde{X}_j^m = \{x_i^m, ..., x_{T'}^m | T^{'} < T\}$, and the reconstructed sequence is $\hat{X}_t^m$, and the aim of this method is to regenerate the masked sequence, the loss is defined as:

$$L_t = \sum_{i=1}^{N} \sum_{t=T^{'}+1}^{T} ||\hat{x}_t^i - x_t^i||_2^2 \tag{7}$$

where N is batch size.

The self-supervised task is trained by minimizing the overall pretrain loss L. And for the task of action recognition, our method initializes the encoder with the weights trained by the self-supervised tasks and uses standard cross-entropy loss to train the adding classifications layer without changing the weight of the encoder.

4. Experiments

To demonstrate the effectiveness of the contrastive-pretext framework, we conduct experiments on three datasets: the NTU RGB+D dataset, the NTU RGB+D 120 dataset and the North-Western UCLA dataset. Our goal is to demonstrate that the features learned by encoder and pretext tasks are well enough for action recognition.

4.1. Datasets and Settings

4.1.1. NTU-RGB+D 60 (NTU-60)

It contains more than 56000 sequences in 60 classes in which three camera views are recorded simultaneously from different perspectives and with 25 joints for each body. We follow the provided evaluation protocol: (a) Cross-Subject (X-Sub) setting separates training and testing set by different persons. (b) Cross-View (X-View) setting covers 37646 samples captured by one camera for training and samples from the other camera are for testing.

4.1.2. NTU-RGB+D 120 (NTU-120)

It is the extension of NTU-60, whose scale reaches 120 classes for action, 106 participants and 113945 sequences. Similarly, there are also two validation protocols: Cross-Subject (X-Sub) and Cross-Setup(X-Set). In X-Sub, samples performed by 53 persons are for training and the others are for testing. In X-Set, all 32 setups are separated as half for training and the other for testing.

4.1.3. North-Western UCLA(NW-UCLA)

This dataset contains 1494 videos in 10 action categories performed by 10 subjects. Each skeleton contains 20 joints. Similar to Lin et al. (2020)we use the first two views for training and the third view for testing, which contains 1, 018 videos and 462 videos, respectively.

4.1.4. Model Setting

To train the network, all the sequences are temporally downsampled to 200 frames. There are 6 layers of st-gcn units as encoder. For pretext tasks, there are 5 layers of st-gcn units for regeneration and for each layer we randomly dropout the features at 0.2 probability to reduce the decoder's ability that may learn better representations according to P&C(Su et al. (2020)). We evaluate feature representations with a linear classifier, which is trained on top of the frozen encoder. The classification accuracy is used as a measurement for the representation learning. Adam optimizer is used to optimize our network during training. The learning rate declines from 0.00005 to 0 for 100 epochs. We trained on NVIDIA M40 GPU with batch size of 256 for NTU-60 and NTU-120, while using batch size of 64 for UCLA dataset.

4.2. Comparison with State-of-the-Art

We compare the accuracy of our method with previous state-of-art methods. To give a comprehensive evaluation, we conduct experiments under different settings, including unsupervised, semi-supervised approaches.

Models	**NW-UCLA**
LongT GAN (Zheng et al. (2018))	74.30
MS^{2}L (Rao et al. (2020); Lin et al. (2020))	76.81
P&C (Su et al. (2020))	84.9
Ours	85.12

Table 1: Comparison of action recognition results in unsupervised setting on NW-UCLA

4.2.1. Unsupervised Learning

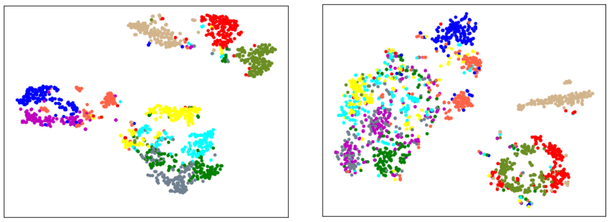
In unsupervised learning, we train the feature extraction $f(\cdot)$, i.e., the encoder in a self-supervision style without any label information, then the feature representation is evaluated by linear classifiers. In our experiments, the linear classifer is trained on top of the encoder. When training the classifier, the encoder is frozen. In Table 1, Table 2 and Table 3, we can see our approach achieve better than all of the unsupervised methods, and even better than some of the supervised methods. This improvement verifies that our methods can extract

Methods	NTU-60	
	X-View	X-Sub
Supervised		
HBRNN (Du et al. (2015b))	64.0	59.1
Deep LSTM (Shahroudy et al. (2016))	67.3	60.7
Part-aware LSTM (Shahroudy et al. (2016))	70.3	62.9
TSA (Caetano et al. (2019))	78.7	70.5
ST-GCN (Yan et al. (2018))	88.3	81.5
AGCN (Shi et al. (2019))	93.7	88.5
Unsupervised		
LongT GAN (Zheng et al. (2018))	48.1	39.1
P&C (Su et al. (2020))	76.1	50.6
MS^{2}L (Lin et al. (2020))	-	52.6
AS-CAL (Rao et al. (2020))	64.8	58.5
CP-STN(Ours)	**76.6**	**69.4**

Table 2: Comparison of action recognition results with supervised and unsupervised learning approaches on NTU-60

Methods	NTU-120	
	X-Sub	X-Set
Supervised		
Part-aware LSTM (Shahroudy et al. (2016))	25.5	26.3
Soft RNN (Hu et al. (2018))	36.3	44.9
TSA (Caetano et al. (2019))	62.9	63.0
Unsupervised		
P&C (Lin et al. (2020))	42.7	41.7
AS-CAL (Rao et al. (2020))	49.7	48.9
CP-STN(Ours)	**55.7**	**54.7**

Table 3: Comparison of action recognition results with supervised and unsupervised learning approaches on NTU-120



(a) t-SNE of CP-STN

(b) t-SNE of P&C

Figure 3: t-SNE visualization of learned features on CP-STN and P&C in the same ten classes (randomly selected) of NTU-60 XSub. Each skeleton sequences is visualized as a point, with skeleton sequences belonging to the same action class having the same color(best viewed in color).

effective features. In Figure 3 a t-SNE visualization of the learned features for the randomly selected 10 classes of NTU-60 XSub testing sets is shown. It is obvious that representations extracted by our proposed method are more semantically separable compared to the P&C. We quantitatively observe that better sequences representations were able to be learned by our proposed method.

4.2.2. Semi-supervised Learning

The proposed CP-STN could be exploited for semi-supervised learning by fine-tuning on a certain fraction of labeled data. The training process utilizes both labeled and unlabeled data. First, we sample data of NTU-120 XSub datasets in a class-balanced way. Then, a linear classifier is attached to the pre-trained network, after which the overall network is jointly fine-tuned with labeled data. Finally, we test the learned representations by freezing the CP-STN model and training the linear classifier with labeled data. Table 4 shows our method's performance, which always outperforms ST-GCN and AS-CAL. We also discover that a purely unsupervised setting for linear evaluation can even surpass using little labels (1% and 5% label fraction). These results identify that the pretrained CP-STN model can learn useful action representations from unlabeled data. After adding insufficient labeled data even reduces the capabilities of the model.

Methods	Label Fraction(%)							
	0	1	5	10	20	30	40	50
Top-1 Accuracy								
ST-GCN (Yan et al. (2018))	-	16.4	34.0	43.1	52.7	59.9	63.9	66.7
AS-CAL (Rao et al. (2020))	48.9	-	-	42.3	-	-	-	52.6
CP-STN(Ours)	**54.7**	**48.8**	**51.5**	**56.1**	**60.6**	**63.2**	**66.5**	**67.8**
Top-5 Accuracy								
ST-GCN (Yan et al. (2018))	-	39.7	63.4	71.3	79.0	83.8	86.0	87.9
AS-CAL (Rao et al. (2020))	78.9	67.1	-	74.3	-	-	-	81.3
CP-STN(Ours)	**81.5**	**77.8**	**80.1**	**82.3**	**85.0**	**86.6**	**87.8**	**89.0**

Table 4: Comparison of action recognition results in unsupervised and semi-supervised setting on NTU-120 XSub

4.3. Ablation Study

We examine the effectiveness of proposed components in the network in this section by action recognition experiments on NTU 60 X-Sub protocols. We pretrain the encoder and then fine-tune the overall network with fixed encoder weight.

4.3.1. Comparison of Different Encoders

We evaluate the necessity of using spatial-temporal graph convolution operation at first. We build baseline framework based on Recurrent Neural Networks (RNN) to model temporal evolution of different actions. For RNN/bilinear LSTM and bilinear GRU, we encode

Encoder	Aug1	Pre1	Aug2	Pre2	Half-C	Acc
Sequence-based						
RNN	S	×	T	×	–	0.4355
BiLSTM	S	×	T	×	–	0.5714
BiGRU	S	×	T	×	–	0.5738
Graph-based						
A+B	S	×	T	×	–	0.4424
A+C	S	×	T	×	–	0.6407
A+B+C	S	×	T	×	–	0.6306
A	S	×	T	×	–	**0.6734**
Graph-based(A)	S	√	T	√	×	0.6846
	S	√	S	√	√	0.6815
	S	√	T	√	√	**0.6937**

Table 5: Performance comparison of CP-STN on NTU-60 X-Sub, in which S refers to spatial augmentation, T refers to temporal augmentation, if half-C is true, two pretexts get half of channels of features extracted from encoder separately.

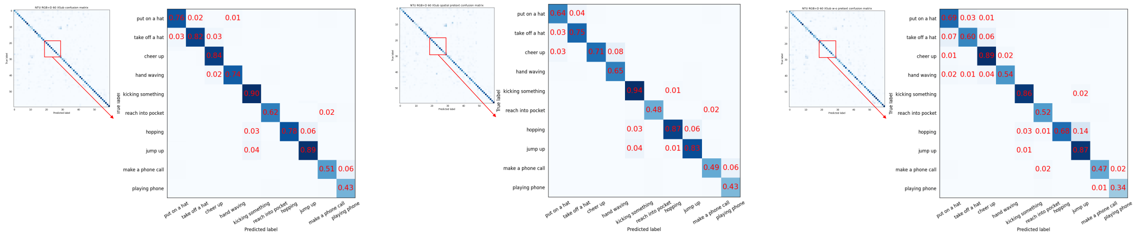
all the joint positions in each frame to a feature vector and feed these vectors into the encoder to learn representations. In Table 5, we can observe that graph-based encoder show outstanding performance.

Compared with only using A to represent the structure of joints connections, learnable joints connections may damage the action classification accuracy in unsupervised representation learning, though it performs well under strong supervision. It mainly because without supervision, the network has limits on joints connection learning. Firstly, without supervision information, unconstrained graph B_k is confused by similar actions. Second, contrastive learning aims to capture invariant information of augmented data, C_k which leverages attention to learn data-dependence relations between different joints helps the model make the graph structure become more similar, which is opposite to the aim of contrastive learning. As a result, we use only A for spatial-temporal information extraction.

4.3.2. Comparison of Different Pretexts

We evaluate different combinations of transformation operators for contrastive learning, i.e., both spatial transformations pretext and spatial-temporal combination pretext. Table 5 shows that adding pretexts is helpful for representation learning. Besides, combine spatial and temporal information in pretext tasks to assist contrastive can achieve the best performance.

Compared to contrastive learning framework, our method combines it with pretext tasks and gathers more fine-grained spatial-temporal information. As shown in Figure 4, after adding spatial and temporal pretexts, the network can judge better on some temporal confusion actions compared with only use spatial pretext tasks and without adding pretext tasks, such as putting on and taking off a hat, which are opposite in the time dimension . The network also performs better on actions with strong continuity in the time dimension, for example, hand waving. Besides, actions such as kicking something and hopping (jump



(a) With spatial temporal pretext tasks

(b) With spatial pretext tasks

(c) Without pretext tasks

Figure 4: Visualization of confusion matrices for testing CP-STN performance with/without different pretext tasks on NTU-60 Xsub. Every item indicates the probability of the cases belonging to row labels that had been classified as the column labels. The higher the probability is the deeper the color is in that item. With deepest color in the diagonal indicates that adding both spatial and temporal pretext tasks can perform best.

up with one foot) that need more joints to cooperate can be verified better after adding spatial pretext task. Furthermore, we also do experiments about channel split strategy to demonstrate it can benefit the the network accentuates the features that are adapted to spatial information or temporal information separately.

5. Conclusion

This paper proposes a novel spatial-temporal enhanced contrastive and pretext learning framework named contrastive-pretext spatial-temporal network (CP-STN) to address the problem of lacking spatial-temporal information extraction in recent unsupervised skeleton-based action representation learning methods. By combing pretext and contrastive learning methods, our method enable network extracting discriminative representations with spatial-temporal information fully. Besides, CP-STN adopts spatial-temporal graph to exploit spatial and temporal connections among adjacent joints effectively. With comprehensive and thorough experiments on two datasets, we can show that our model is a powerful feature extractor that significantly outperforms recent methods.

Acknowledgments

This work was supported in part by the National Key RD Program of China 2020YFB1807800, National Natural Science Foundation of China under Grants (6217105762071067, 62101064, 62001054 and 61771068), Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), National Postdoctoral Program for Innovative Talents under Grant BX20200067, China Postdoctoral Science Foundation under Grant 2021M690469Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583. IEEE, 2015a.

Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015b.

Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang. Early action prediction by soft regression. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2568–2583, 2018.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.

Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1623–1631. IEEE, 2017.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3595–3603, 2019.

Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020.

Jianbo Liu, Yongcheng Liu, Ying Wang, Véronique Prinet, Shiming Xiang, and Chunhong Pan. Decoupled representation learning for skeleton-based gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5751–5760, 2020a.

Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020b.

Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *arXiv preprint arXiv:2008.00188*, 2020.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019.

Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1227–1236, 2019.

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *arXiv preprint arXiv:1611.06067*, 2016.

Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. doi: 10.1109/CVPR.2014.339.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018.

Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.