Unsupervised Feature Enrichment and Fidelity Preservation Learning Framework for Skeleton-based Action Recognition

Chuankun Li, Shuai Li, *Senior Member, IEEE,* Yanbo Gao, Xingyu Gao, Ping Chen, Jian Li
Wanqing Li, *Senior Member, IEEE*

***Abstract*—Unsupervised skeleton-based action recognition has achieved remarkable progress recently. Existing unsupervised learning methods suffer from severe overfitting problem, and thus small networks are used, significantly reducing the representation capability. To address this problem, the overfitting mechanism behind the unsupervised learning for skeleton-based action recognition is first investigated. It is observed that skeleton is already a relatively high-level and low-dimension feature, but not in the same manifold as the features for action recognition. Simply applying the existing unsupervised learning method tends to produce features that discriminate the different samples rather than action classes, resulting in the overfitting problem. To address this problem, this paper proposes an Unsupervised spatial-temporal Feature Enrichment and Fidelity Preservation (U-FEFP) learning framework to generate rich distributed features that contain all the information of a skeleton sample. A spatial-temporal feature transformation subnetwork is developed using channel-wise topology refinement graph convolutional block and graph convolutional gated recurrent unit block as the basic feature extraction network. The unsupervised Bootstrap Your Own Latent-based learning is utilized to generate rich distributed features, and the unsupervised pretext task-based learning is employed to preserve the information contained in the skeleton. The two unsupervised learning ways are collaborated as U-FEFP to produce robust and discriminative representations. Experimental results on four widely used benchmarks, namely NTU-RGB+D-60, PKU-MMD, NTU-RGB+D-120 and UAV-Human dataset, demonstrate that the proposed U-FEFP obtains the best result compared with the state-of-the-art unsupervised learning methods.**

***Index Terms*—Skeleton, Action recognition, Graph convolutional network, Unsupervised learning**

I. Introduction

ACTION recognition using different modalities (e.g., video, skeleton) [1–7] has been actively studied due to its use in many practical applications such as video surveillance and behaviour analysis. Compared with the RGB video, 3D skeleton as a high-level representation is light-weight and robust to both view differences and intricate background. Therefore,3D skeleton-based action recognition has been widely investigated, including methods based on handcrafted features [8, 9], Convolutional Neural Networks (CNNs) [10–13], Recurrent Neural Networks (RNNs) [14–17] and Graph Convolutional Networks (GCNs) [18–21]. However, these methods are developed in a supervised manner and require extensive annotated labels, which is inconvenient. Learning general features from unlabelled data for 3D skeleton-based action recognition is still challenging and highly desired.

There are two main approaches for unsupervised skeleton-based action recognition. The first approach utilizes an encoder-decoder network and generates useful features by pretext tasks such as auto-regression [22], reconstruction [23] and jigsaw puzzle [24]. Existing methods [22–24] usually take advantage of the RNNs to encode the input skeleton sequence and then regressively predict them. Their performance on the downstream task is highly dependent on the design of pretext tasks. The second approach utilizes the contrastive learning such as Bootstrap Your Own Latent (BYOL) [25], Momentum contrast [26]. These methods [27–30] learn features by pulling or pushing the features of different samples as positive and negative pairs. Putting aside their individual problems such as designing relevant task and differentiating positive and negative pairs, both approaches suffer from severe overfitting. The existing networks used in supervised learning [19–21] cannot work effectively in the unsupervised learning due to this severe overfitting. Consequently, the existing unsupervised learning methods employ very simple models, either using only basic RNN models [31–34] or using very small models with fewer neurons [35–41]. However, such simple models with low-dimension features are not capable of the high-level action recognition task, and thus cannot achieve high performance.

Currently, there is no investigation on the mechanism behind the severe overfitting problem in the unsupervised learning for skeleton-based action recognition. This paper first studies the overfitting mechanism in the unsupervised

Manuscript received Sep. 10, 2024.
This work was supported in part by the National Natural Science Foundation of China (No. 62101512, 62001429, 62271453 and 62271290) and Shanxi Scholarship Council of China (2023-131). (Corresponding author: Shuai Li, Xingyu Gao)

C. Li, P. Chen and J. Li are with the State Key Laboratory of Dynamic Testing Technology and School of Information and Communication Engineering, North University of China, Taiyuan 030051, China; (e-mail: chuankun@nuc.edu.cn; chenping@nuc.edu.cn; lijian@nuc.edu.cn)
S. Li and Y. Gao are with School of Control Science and Engineering and School of Software, respectively, Shandong University, Jinan 250100, China; (e-mail: shuaili@sdu.edu.cn; ybgao@sdu.edu.cn)
X. Gao is with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China (e-mail: gxy9910@gmail.com).
W. Li is with the Advanced Multimedia Research Lab, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: wanqing@uow.edu.au).

skeleton-based action recognition learning and shows that the existing unsupervised learning method cannot effectively generate features that are highly relevant and useful for action recognition. With skeleton sequences already being relatively high-level and low-dimension representations, the encoder-decoder architecture and the contrastive learning can easily generate features representing or differentiating each skeleton sequence, and such features may not be useful for action recognition. This can be intuitively understood since the high-level skeleton is not in the same manifold as the high-level features for action recognition (considering the example that directly using one fully connected layer cannot produce high action recognition performance). This is further illustrated in the following Motivation Section.

To tackle the above problem, we present an Unsupervised spatial-temporal Feature Enrichment and Fidelity Preservation framework (U-FEFP). The proposed network generates *rich distributed spatial-temporal* features *containing all information of the original skeleton*. Our contributions can be summarized as follows.

- We investigate the mechanism behind the severe overfitting problem in the unsupervised learning for skeleton-based action recognition. It is found that features representing each skeleton may not be aligned with the features for action recognition, leading to the requirement of learning rich distributed features. To the best of our knowledge, this is the first research that investigates the overfitting mechanism in the unsupervised skeleton-based action recognition.
- Based on our observation on the overfitting mechanism, an Unsupervised spatial-temporal Feature Enrichment and Fidelity Preservation (U-FEFP) learning framework is developed by taking advantage of the BYOL-based learning and pretext task based learning. It learns rich distributed features while preserving the fidelity of the original skeleton to solve the overfitting problem.
- A spatial-temporal feature transformation subnetwork is developed by combining the channel-wise topology refinement graph convolution network (CTR-GCN) and the graph convolutional GRU network (GConv-GRU).

Extensive experiments verify the capacity of the representations learned by our U-FEFP, on NTU-RGB+D-60 [42], PKU-MMD [43], NTU-RGB+D-120 [44] and UAV-Human [45] datasets. It achieves the best results under both the unsupervised and semi-supervised training.

II. Related works

This section briefly reviewed the works related to the proposed method, including supervised skeleton-based action recognition and unsupervised skeleton-based action recognition [8, 10–13, 18, 24, 27, 46–48].

A. Supervised Skeleton-based Action Recognition

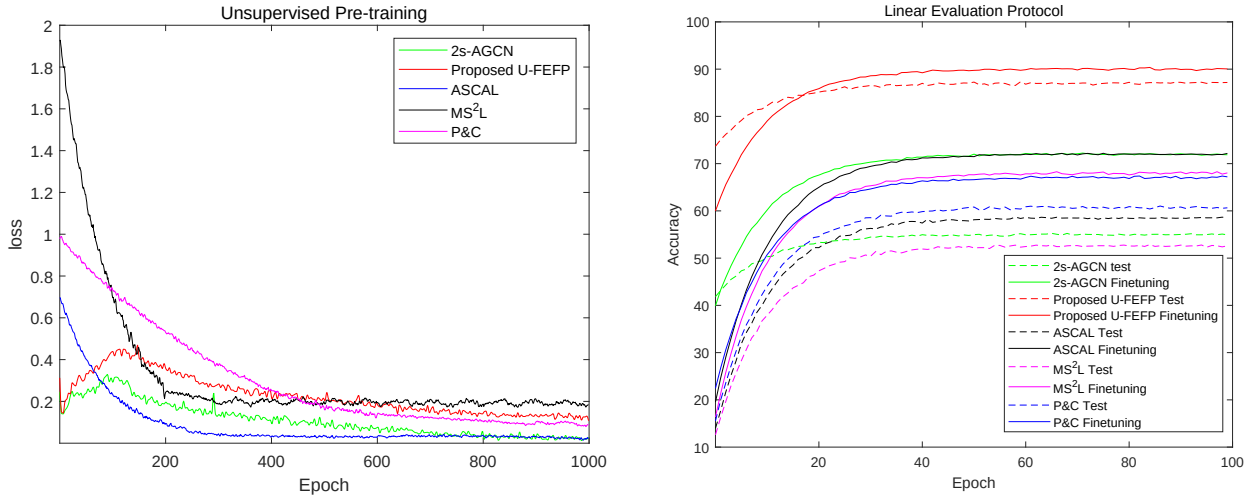
1) Conventional Hand-Crafted Feature-based Method: The hand-crafted skeleton features are widely used in early action recognition [8, 9, 49–51]. For example, Weng et al. [8] used Naive-Bayes Nearest-Neighbor to capture the structure of skeleton joints in the spatio-temporal dimensions. However, the hand-crafted skeleton features are usually hard to be generalized and these methods perform worse on large datasets such as NTU-RGB+D-120 [44] datasets.

2) Deep Learning-based Method: Depending on the type of network, it can be generally categorized into three approaches: CNNs-based, RNNs-based and GCNs-based. In the category of CNNs based methods [10–13, 52–56], a skeleton sequence is mapped to a color image and then recognized into action classes with CNNs. For example, Hou et al. [12] painted skeleton joints with different colors to generate skeleton optical spectra image. Banerjee et al. [53] used distance feature, distance velocity feature, angle feature and angle velocity feature to obtain four grayscale images. The fuzzy combination is used to fuse scores extracted from four grayscale images. Xia et al. [54] utilized convolutions with attention mechanisms to generate local-and-global attention network. Zhu et al. [55] designed an attention mechanism with using cuboid CNN model, where a cuboid arranging strategy is used to organize new action representation among all body joints. Although the temporal information is explored to some extent by coding the temporal change into an image, its representation capability in temporal modelling is still relatively limited.

The second category is to treat skeleton as a sequence and utilizes RNNs to obtain spatial-temporal feature. It focuses more on the temporal feature while the spatial feature of skeleton joint is not fully explored. To enhance the capturing of spatial information, many methods [14–17, 57–59] have been developed. For example, a denoising sparse LSTM [57] is proposed to decrease the intra-class diversity and extract more spatial-temporal information. Ng et al. [59] proposed the multi-localized sensitive autoencoder-attention-LSTM to reduce negative variations such as performers and viewpoints and improve performance. Zhang et al. [58] selected a lot of simple geometric features to feed into a several RNN architecture with a new smoothed score fusion method to improve recognition accuracy. However, these approaches cannot effectively capture relationship among joints.

In order to solve this problem, the third approach uses GCNs [18–21, 60–62] to capture topological graph structure of skeleton. For example, spatial-temporal graph convolutional network (ST-GCN) [18] is proposed to obtain topological spatial-temporal information, where a static graph is used to capture relationship among joints. However, a static graph is not suitable for all different actions and cannot extract dynamic features among spatial joints. In order to solve this problem, existing methods adaptively learned the topology with attention or other similar methods. For example, Liu et al. [62] proposed a Graph Convolutional Networks-Hidden conditional Random Field (GCN-HCRF) model to construct multi-stream framework. Generally, GCNs-based methods have achieved good result in the supervised skeleton-based action recognition.

While the supervised learning-based methods have greatly advanced in the last few years and achieved good performance, these methods require massive labels for training and cannot effectively work for unlabeled skeleton data. Therefore, unsu-



(a) (b)

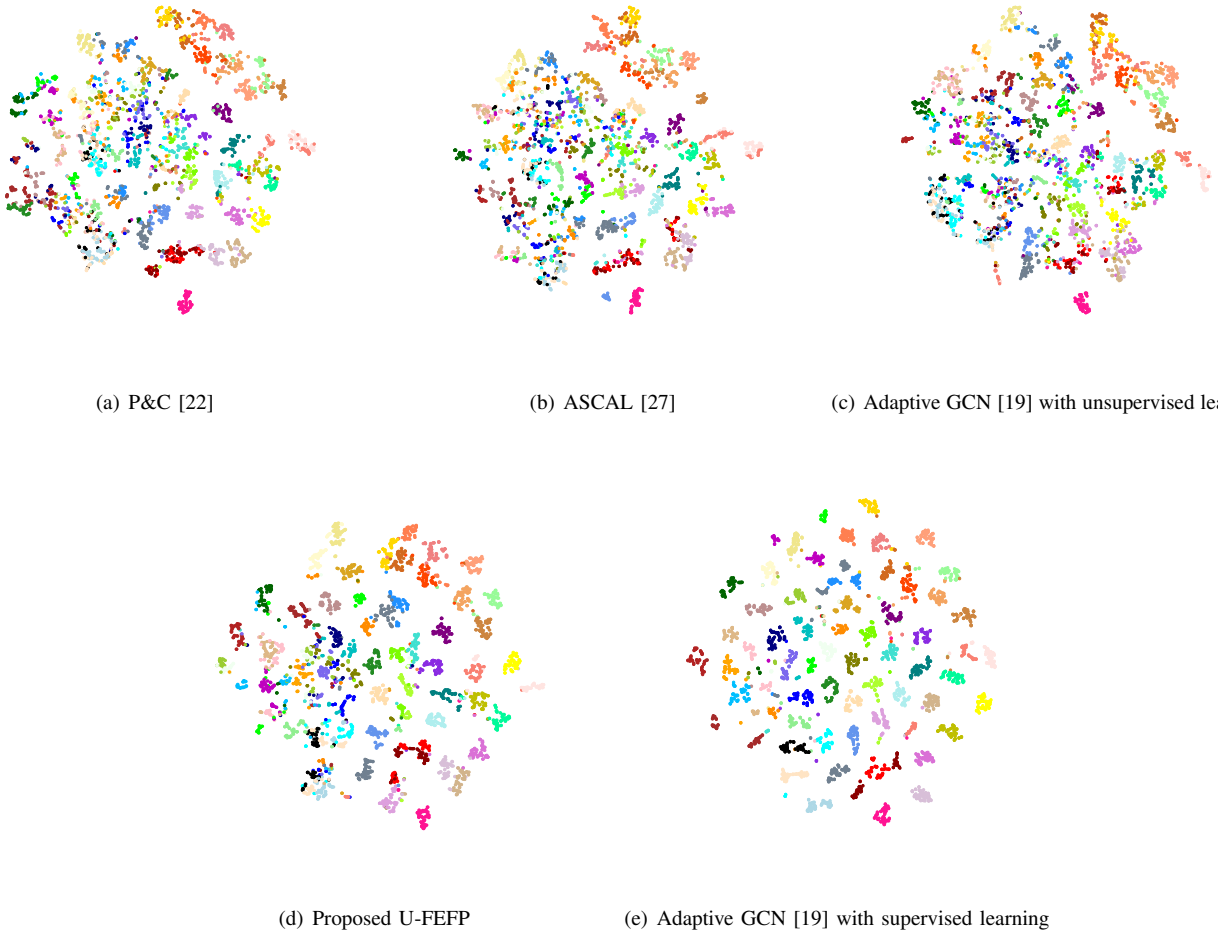
Fig. 1: Illustration of the overfitting problem in the unsupervised learning for skeleton-based action recognition. (a) Comparisons of the proposed U-FEFP and the existing methods in terms of training loss in the unsupervised pre-training, and (b) Comparisons of the accuracy in the fine-tuning/test. The smaller training loss achieved by some existing methods such as 2s-AGCN [19] with much worse test performance over the proposed method demonstrates the overfitting problem. The performance gap between the finetuning and test results also indicates the overfitting problem.

pervised skeleton-based action recognition methods are highly desired.

B. Unsupervised Skeleton-based Action Recognition

1) Self-supervised learning-based Method: Pretext tasks are designed to extract discriminative features in self-supervised learning-based methods. The encoder-decoder model and the Generative Adversarial Network (GAN) [23] are used to reconstruct the skeleton sequence. Su et al. [22] designed an autoencoder structure with a weak decoder using recurrent neural network to learn more robust features from skeleton sequence. Lin et al. [24] proposed two pretext tasks including motion prediction and Jigsaw puzzle recognition to learn more general representations. However these methods usually used RNNs to extract temporal features where the spatial information is not mined effectively. So many works [63–65] based on Transformer have been proposed to improve performance. Cheng et al [63] designed a Hierarchical Transformer to predict the motion between adjacent frames. Kim et al. [64] utilized global and local attention mechanism to predict multi-interval pose displacement. Mao et al. [65] proposed a Masked Motion Prediction (MAMP) framework using Transformer in encoder-decoder.

2) Contrastive Learning-based Method: Rao et al. [27] used momentum LSTM with a dynamic updated memory bank, and augmented instances of the skeleton sequence were used to learn feature representation in a contrasted way. A cross-view contrastive learning scheme [28] is designed and leveraged multi-view complementary signal for supervision. Several skeleton-specific spatial and temporal augmentations [29] have been designed to construct skeleton intra-inter contrastive learning. Lin et al. [66] proposed a new actionlet dependent contrastive learning by treating motion and static regions differently. Zeng et al. [67] proposed a Cross Momentum Contrast (CrossMoCo) framework to learn local and global semantic features and used two independent negative memory banks to improve high-quality of negative samples. Gao et al. [68] proposed spatio-temporal contrastive learning using different spatio-temporal observation scenes to build contrastive proxy tasks. Shah et al. [32] proposed Hallucinate Latent Positives for contrastive learning to generate new positives and improve performance. These methods need to design different positive and negative pairs for better learning. A skeleton-based relation consistency learning scheme is developed to expand the contrastive objects from individual instance to the relation distribution between instances, and target at pursuing the relationship consistency learning between different instances. Zhang et al. [36] used Barlow Twins' objective function to minimize the redundancy and keep similarity of different skeleton augmentations. However, these methods cannot effectively capture robust and discriminative features for action recognition. Moreover, both the self-supervised learning-based and contrastive learning-based methods suffer from severe overfitting problem and only very small networks can be used, leading to reduced representation capability. Lin et al. [34] used equivariant contrastive learning to obtain a more discriminative representation space. Zhang et al. [33] proposed prompted contrast with masked motion modeling and used the contrastive learning and masked prediction tasks to improve performance. Shu et al. [69] designed multi-granularity anchor-contrastive representation learning to obtain multi-granularity action representations. Zhou et al. [39] used a partial spatio-temporal learning (PSTL) framework and designed spatio-temporal mask to improve performance. Zhu et al. [37] proposed a relative visual tempo contrastive learning framework to capture motion features.



(a) P&C [22] (b) ASCAL [27] (c) Adaptive GCN [19] with unsupervised learning

(d) Proposed U-FEFP (e) Adaptive GCN [19] with supervised learning

Fig. 2: t-SNE visualization of the learned features of different methods on the cross-subject of NTU-RGB+D-60. 60 samples are selected for each class on the dataset. (a) Unsupervised learning-based on pretext task, P&C [22]. (b) Unsupervised contrastive learning with the momentum LSTM, ASCAL [27]. (c) Unsupervised contrastive learning with the adaptive GCN [19]. (d) Proposed U-FEFP. (e) Supervised learning with the adaptive GCN [19].

III. Motivation

As mentioned in the Introduction and Related Work sections, existing unsupervised learning for skeleton-based action recognition methods suffer from severe overfitting problem. This is quite different from the image related unsupervised learning, where the losses in the training aligns with the test accuracy, i.e., smaller training loss leading to higher test accuracy. To better illustrate the overfitting problem in the unsupervised learning for skeleton-based action recognition, the unsupervised pre-training, fine-tuning and test processes of the existing 2s-AGCN [19], AS-CAL [27], MS^{2}L [24], P&C [22] and the proposed U-FEFP are visualized in Fig. 1. Detailed descriptions on the experimental setup are shown in the following Subsection V-C. As shown in Fig. 1, while the existing methods such as 2s-AGCN [19] achieve smaller loss than the proposed U-FEFP in the training, the test accuracy is much worse than the proposed one, demonstrating its overfitting in the unsupervised learning process. Moreover, the performance gap between the finetuning and test results is much larger for the existing methods than ours, further indicating the overfitting problem. To solve this problem, we study why one model working well in supervised learning leads to overfitting in unsupervised learning for skeleton-based action recognition, and this behaviour has not occurred in the image related unsupervised learning. Instead of directly reducing the network parameters as in [35, 36] which in turn reduces its representation capability, this paper first investigates the mechanism behind this severe overfitting problem. Then based on the observation of this overfitting mechanism, our learning framework, U-FEFP, is proposed.

To illustrate the differences between the features learned with unsupervised learning and supervised learning, t-SNE [70] is used to visualize their embedding clustering. The visual illustration shows how the features of the same class of actions form clusters while different classes of actions are separated. Features from three unsupervised learning methods, including unsupervised learning based on pretext task (P&C [22]), unsupervised contrastive learning with the momentum LSTM (ASCAL [27]), unsupervised contrastive

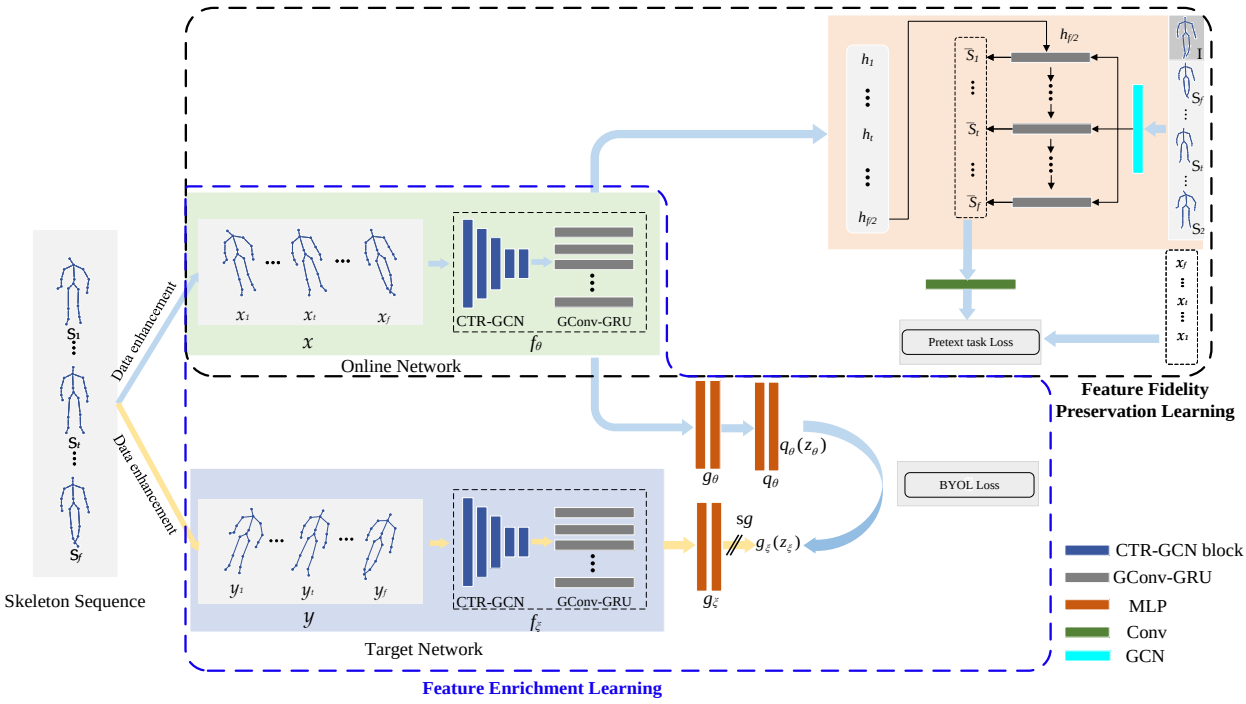


Fig. 3: The framework of the proposed U-FEFP, with unsupervised BYOL-based feature enrichment learning and unsupervised pretext task based fidelity preservation learning. It consists of an online network (in green), a target network (in blue) and a reversed prediction network (in beige). The online network is trained to learn rich representations and the target network is slowly updated by the exponential moving average of the online network to make them asynchronous. The reversed prediction network is used to reconstruct the skeleton sequence with the features generated by the online network. The BYOL-based contrastive learning (within the black dash box) and the reversed prediction (pretext task)-based learning (within the red dash box) are used to keep similarity of different skeleton augmentations at feature and instance level, respectively.

learning with the adaptive GCN [19], are illustrated, and features from the supervised learning with the adaptive GCN [19] are used for comparison. The t-SNE comparison is shown in Fig. 2 using 60 samples from each action class. First, by visualizing the t-SNE illustration of the supervised GCN in Fig. 2(e), it can be seen that the samples are clustered well according to different actions, leading to the good classification results with supervised learning. On the contrary, the t-SNE illustrations of the unsupervised learning with different methods in Figs. 2(a), 2(b) and 2(c) show that the samples are also grouped to some extent, but not according to their action classes, thus producing poor results. Especially comparing the t-SNE illustrations of the adaptive GCN under supervised and unsupervised learning in Figs. 2(e) and 2(c), it can be clearly seen that with the same network, the features are learned completely differently, in terms of their clustering behaviour to the action classes. Intuitively, this can be analyzed as the choice of the negative samples not highly related to action recognition, thus generating features not clustered as action classes.

As a matter of fact, the skeleton sequences are already relatively high-level and low-dimension representations. In such a case, unsupervised learning tends to produce features that directly discriminate or reconstruct samples, and such features may not be useful for action recognition. In other words, the features learned in the unsupervised way distribute in a high-level manifold that is not aligned with the high-level feature manifold of the action recognition. Accordingly, the loss in the unsupervised learning can be very small while the loss of the action recognition is very high, leading to the overfitting problem. To overcome this problem, we propose a U-FEFP learning framework to generate *rich distributed spatial-temporal* features *containing all information of the original skeleton* in unsupervised learning. The *rich distributed spatial-temporal* features contain distributed features that can be useful for action recognition, instead of pushing features to discriminate certain samples that may narrow the representation capability of the features. Constraining the features to be *containing all information of the original skeleton* encourages the network to preserve all useful information. To the best of our knowledge, this is the first research that clearly points out to learn such features in the unsupervised skeleton-based action recognition. The t-SNE illustration of our U-FEFP learning framework is shown in Fig. 2(d). Compared to the other unsupervised learning methods shown in Figs. 2(a), 2(b) and 2(c), our U-FEFP clearly produces features that are better aligned with the action classes. Although certain samples may deviate from the action centers, they are also away from other action centers, making them easier for recognition.

IV. Proposed Method

The framework of the proposed unsupervised spatial-temporal feature enrichment and fidelity preservation network (U-FEFP) is shown in Fig. 3, consisting of two parts: unsupervised feature enrichment learning based on BYOL and unsupervised fidelity preservation learning based on pretext task. The unsupervised feature enrichment learning, including the online network and target network in BYOL, is developed for spatial-temporal feature transformation to obtain a rich distributed spatial-temporal feature representation. The unsupervised fidelity preservation learning, including the online network and the decoding network in the pretext task, is developed to keep the information of original skeleton. The details of the proposed U-FEFP are presented in the following.

A. Unsupervised Spatial-temporal Feature Enrichment Learning

1) Spatial-temporal Feature Transformation Network: A channel-wise topology refinement graph convolution (CTR-GCN) followed by a graph convolutional GRU network (GConv-GRU) is developed as the basic architecture of the online network and target network used in our unsupervised BYOL-based learning, to produce the spatial-temporal features. CTR-GCN takes advantage of the graph convolution to extract the skeleton feature in the spatial-temporal dimension. With the expressive power of graph convolution in processing non-grid data like skeleton, it can obtain rich spatial features. Moreover, considering the temporal change of each joint among frames is also important in characterizing the spatial features of a skeleton to be representative and discriminative against others, CTR-GCN is used to obtain spatial and short-term temporal features. In each CTR-GCN block, it contains one spatial graph convolution extracting the spatial information and one temporal convolution mining the short temporal information. The basic structure of the graph convolution is the same as [71], which is not further detailed here. Five CTR-GCN blocks are used and the numbers of convolution kernels are 64, 64, 128, 256 and 512, respectively. In order to reduce the computation, in the second and fourth block, the stride of temporal convolution is set to 2, which halves the length of the temporal features.

For capturing the long-term temporal features, a GConv-GRU network is used, which aggregates the spatial and short-term temporal features from CTR-GCN to obtain long-term information. On one hand, it is found that using CTR-GCN for complete spatial-temporal feature extraction is easily overfitting, since multiple CTR-GCN layers are needed to extract the long-term temporal features, making the network complex. By contrast, in our method, as shown in Fig. 3, only five CTR-GCN blocks (versus ten blocks in the conventional CTR-GCN methods [71]) are used to reduce overfitting. On the other hand, GConv-GRU, due to its recurrent structure, is more suitable for decoding features in the following temporal pretext task (described in the next subsection). Therefore, a sequential architecture combining the CTR-GCN and GConv-GRU is used in this paper.

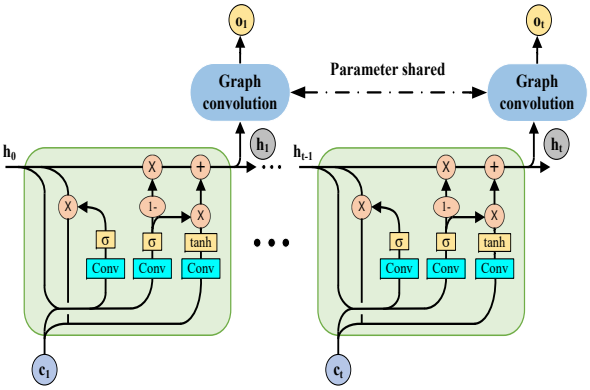


Fig. 4: The structure of the GConv-GRU

For the GConv-GRU, while the recurrent structure aggregates the temporal information, its sequential processing also incurs great computation if the processing of each step is computationally expensive. Here, considering the features are captured via the CTR-GCN with graph convolution, the spatial structure information is already contained in the input features. Therefore, in order to reduce computation, general convolution, with 1*1 kernel for per-joint processing, is used in the recurrent update of each GRU step. To make the input processing consistent with the recurrent processing, general convolution is also used in the input processing. The hidden output of the GConv-GRU is further enhanced with graph convolution, which can be computed in parallel over all time steps with less computation complexity and enhancing the temporal features with the spatial structure. The update of the GConv-GRU is shown in Fig. 4.

2) BYOL-based Feature Enrichment Learning: As mentioned in the Motivation, rich distributed features are highly desired for unsupervised learning. It is required to produce a rich set of distributed high-level representation features that are useful to discriminate different samples and useful for the downstream high-level task, i.e., action recognition. Naively we can generate a set of features with a network of random parameters. However, this cannot provide view-invariant (shift-invariant, pose-invariant, etc.) high-level features. Thanks to the BYOL [25]-based feature learning, rich distributed features can be learned with two asymmetric networks, i.e., an online network and a target network as shown in Fig. 3, by pulling together the features of different augmented versions of one sample.

As shown in the Fig. 3, data augmentation is first used to enhance a skeleton sequence to different views. Suppose that an original skeleton sequence $\mathbf{S} = (\mathbf{S_1}, \cdots, \mathbf{S_f})$ contains f consecutive skeleton frames, where $\mathbf{S_i} \in \mathbb{R}^{\mathbf{N}\times\mathbf{3}}$ is 3D coordinates of N skeleton joints. The data augmentation strategy in [29] (i.e., spatial augmentation and temporal augmentation) and rotation are used to transform $\mathbf{S}$ into its augmented versions.The two different views of the samples are then processed by the online network and target network, generating the spatial-temporal features. Then two nonlinear projectors g_θ and g_ξ are used to project the hidden features to a new feature space. Two

nonlinear projectors use same structure and are updated in the same way with online network and target network. The nonlinear projector contains two fully connected layers. The first one is of 1024 neurons followed by batch normalization and activation (relu). The second one contains 512 neurons generating features without the normalization and activation. This up-projects the features back to 512 channels to enrich the representation.

As in BYOL framework, asymmetric architecture is used and a predictor q_θ using the same network as the nonlinear projector g_θ is added only to the online branch to generate prediction $\mathbf{q}_\theta(\mathbf{z}_\theta)$, where $\mathbf{z}_\theta$ is output of the projector g_θ. For the target branch, the stop-gradient operation is used after the nonlinear projector $\mathbf{g}_\xi$ and obtains feature $\mathbf{g}_\xi(\mathbf{z}_\xi)$, where $\mathbf{z}_\xi$ is output of the target branch. Then $\mathbf{q}_\theta(\mathbf{z}_\theta)$ and $\mathbf{g}_\xi(\mathbf{z}_\xi)$ are normalized with the ℓ_2-norm separately as

$$\bar{\mathbf{q}}_\theta(\mathbf{z}_\theta) \triangleq \mathbf{q}_\theta(\mathbf{z}_\theta)/\|\mathbf{q}_\theta(\mathbf{z}_\theta)\|_\mathbf{2} \tag{1}$$

$$\bar{\mathbf{g}}_\xi(\mathbf{z}_\xi) \triangleq \mathbf{g}_\xi(\mathbf{z}_\xi)/\|\mathbf{g}_\xi(\mathbf{z}_\xi)\|_\mathbf{2} \tag{2}$$

Finally, Mean Square Error (MSE) objective function between $\bar{\mathbf{q}}_\theta(\mathbf{z}_\theta)$ and $\bar{\mathbf{g}}_\xi(\mathbf{z}_\xi)$ is used to construct the self-supervised loss and can be expressed as:

$$L_{\theta,\xi} = \left\|\bar{\mathbf{q}}_\theta(\mathbf{z}_\theta) - \bar{\mathbf{g}}_\xi(\mathbf{z}_\xi)\right\|_2^2 \tag{3}$$

which can be further transformed by substituting $\bar{\mathbf{q}}_\theta(\mathbf{z}_\theta)$ and $\bar{\mathbf{g}}_\xi(\mathbf{z}_\xi)$ with Eqs. (1) and (2), respectively, to the following:

$$L_{\theta,\xi} = 2 - 2 \cdot \frac{\langle \mathbf{q}_\theta(\mathbf{z}_\theta), \mathbf{g}_\xi(\mathbf{z}_\xi)\rangle}{\|\mathbf{q}_\theta(\mathbf{z}_\theta)\|_2 \cdot \|\mathbf{g}_\xi(\mathbf{z}_\xi)\|_2} \tag{4}$$

The loss is symmetrized and a symmetric loss $L_{\theta,\xi}^{'}$ can be obtained by feeding the $\mathbf{x}$ and $\mathbf{y}$ into target network and online network, respectively. Finally, the learning loss is obtained as $L_{BYOL} = L_{\theta,\xi} + L_{\theta,\xi}^{'}$.

In the training process, weights ξ of target network are updated using the exponential moving average of the online network weight θ which follows $\tau\xi + (1-\tau)\theta \rightarrow \xi$. This allows the online network and target network to be always asymmetric. The online and target GConv-GRU network produce a temporal feature with half the time steps of the sequence. While the features at all time steps can be processed as above, in this paper for simplicity, a global pooling over the temporal dimension is used to generate the final feature and then processed.

This BYOL-based feature learning enables the online network to generate rich distributed high-level features. Simply speaking (as an extreme case for intuitive understanding), the target network produces a rich set of randomly combined features, while the learning updates the online network and target network to produce view-invariant high-level features. The asymmetric updating avoids them to generate the trivial solutions of zero or other fixed representations. Therefore, the online network produces rich distributed high-level features, which is further constrained by the pretext task (described in the following subsection) and used for the final action recognition.

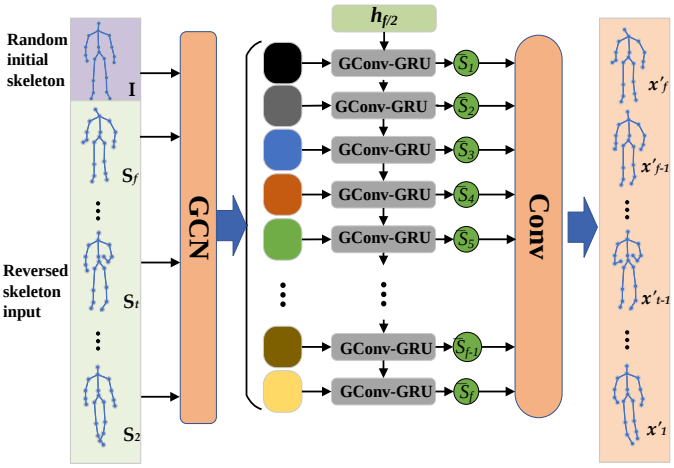


Fig. 5: The structure of the decoder in the unsupervised pretext task-based learning.

B. Unsupervised Spatial-temporal Feature Fidelity Preservation Learning

While the above BYOL-based learning generates rich distributed features that can keep the similarity of augmented different-view skeleton data, it cannot ensure the generated features to be able to classify all actions. In other words, the representation space of the generated features may be reduced since there is no constraint in discriminating different samples or actions. Therefore, to generate features that are not only rich distributed but also full and contain as much information of the original skeleton as possible, a fidelity preservation constraint is required.

In this paper, an unsupervised pretext task-based learning is designed using reversed prediction. Motivated by the pretext task using encoder-decoder network [22, 23], the feature of the online network is used as the encoder feature, and a decoder is used to predict the skeleton sequence. To be specific, the skeleton sequence in the reverse order is used with each skeletal joint predicted. In this way, the hidden state obtained from the encoder (which is at the time step of the last skeleton) can be directly used for predicting the skeleton at its corresponding time step. The decoder used in this paper consists of one GCN, one GConv-GRU and one Convolution as presented in Fig. 5. At each time step, the previous skeleton (also in the reverse order) is first fed into GCN to generate features $\mathbf{d}$ as input, and with the recurrent hidden feature from GConv-GRU at the previous step, the network predicts the skeleton. For the first time step, the encoder (online network) feature is used as the initial recurrent hidden feature and a randomly initialized skeleton is used as the input. The update process can be expressed as

$$(\bar{\mathbf{h}}_\mathbf{t}, \bar{\mathbf{S}}_\mathbf{t}) = \begin{cases} \Theta(\mathbf{h_{f/2}}, \mathbf{d_1}) & t = 1 \\ \Theta(\bar{\mathbf{h}}_{\mathbf{t-1}}, \mathbf{d_{t-1}}) & t > 1 \end{cases} \tag{5}$$

where $\Theta(\cdot)$ denotes the decoder GConv-GRU, and $\mathbf{h_{f/2}}$ is the encoder feature from the online network, i.e., the hidden state at the last step. When initializing the decoding (t=1), $\mathbf{h_{f/2}}$ and randomly initialized $\mathbf{d_1}$ is provided to the decoder

GConv-GRU, producing the recurrent hidden features $\overline{\mathbf{h}}_\mathbf{1}$, and the output feature $\overline{\mathbf{S}}_\mathbf{1}$. Then $\overline{\mathbf{S}}_\mathbf{1}$ is processed with convolution to produce the first skeleton (in the reverse order). For the following time steps, $\mathbf{d_{t-1}}$, the feature of the skeleton at the previous time step, is used as the input and $\overline{\mathbf{h}}_{\mathbf{t-1}}$ is used as the recurrent input. Finally, MSE between the output sequence $\mathbf{x}'$ and $\overline{\mathbf{x}}$ (reversed sequence of $\mathbf{x}$) is used as loss of the pretext task-based unsupervised learning, $L_P = \|\mathbf{x}' - \overline{\mathbf{x}}\|_2^2$.

This unsupervised pretext task-based learning enforces the features generated from the online network to contain all the information of the skeleton so as to predict the original skeleton. Together with the BYOL-based learning, the proposed U-FEFP learning framework is jointly trained using the total loss $L = L_{BYOL} + L_P$, to generate rich distributed and fidelity preserved features.

V. Experimental Results

A. Datasets

Three widely used datasets, including the NTU RGB+D-60 dataset, NTU-RGB+D-120 dataset and PKU-MMD dataset, are used for evaluating the proposed method.

NTU RGB+D-60 dataset (NTU-60): NTU-60 [42] is a commonly used action recognition dataset captured by three Microsoft Kinect v2. It consists of 56880 skeleton sequences captured by using 40 persons and contains 60 actions. Each action contains more than 900 samples. This dataset is captured using three different views including 45 degree view, side view and front view. The test settings suggested along with the dataset [42] are used in this paper, including the cross-subject (X-sub) and the cross-view (X-view).

NTU-RGB+D-120 dataset (NTU-120): NTU-120 [44] further extends NTU-60. It consists of 114480 action clips captured by using 106 unique human subjects with different ages (from 10 to 57). Compared to NTU-60, camera viewpoints are extended to 155 in this dataset. The subjects coming from 15 countries with different cultural backgrounds bring a very realistic variation in the quality of actions. We use cross-setup (X-Set) and cross-subject (X-Sub) adopted in [44] to evaluate the proposed method.

PKU-MMD dataset: PKU-MMD dataset [43] has nearly 20000 action clips and consists of 51 action categories. We utilize two subsets namely, PKU-MMD I and PKU-MMD II. Compared to PKU-MMD I, PKU-MMD II contains more low-quality samples. The cross subject (X-Sub) setting are utilized for both subsets.

UAV-Human: UAV-Human [45] is a relatively smaller benchmark that is captured by unmanned aerial vehicles (UAV) and involves 119 subjects. This dataset encompasses 155 activity classes, such as daily activities, productive activities, violent activities, social interaction behaviors, life-saving activities and UAV control gestures. The dataset is challenging because of motion blurs and the continuously changing attitudes and heights of the UAVs during flight. We follow the paper [45] and adopt the cross-subject evaluation protocol to set the training data and the testing data.

TABLE I: Comparison of different feature transformation networks as online network

Method	X-Sub (%)	X-View (%)
MS-G3D [73]+ BYOL	54.16	57.98
CTR-GCN [71]+ BYOL	54.24	58.62
2s-AGCN [19] + BYOL	56.63	59.29
ST-GCN-v1 [18]+ BYOL	80.22	84.50
ST-GCN-v2 [18]+ BYOL	83.42	87.30
online network v1 + BYOL	64.78	70.36
online network v2 + BYOL	78.35	83.28
online network v3 + BYOL	72.22	77.18
online network v4 + BYOL	83.12	87.80
online network v5 + BYOL	83.62	88.10
Proposed online network + BYOL	**84.85**	**88.80**

B. Implementation Details

Unsupervised Pre-training: The PyTorch framework is used to implement the proposed U-FEFP and run it on four Tesla A100 GPUs. We utilize LARS [35] as optimizer and train 1000 epochs for all datasets. The learning rate grows from 0 to 2 in the first 20 epochs and then decreased to 0.001 based on a cosine decay schedule. We follow paper [72] to downsample 64 frames for each skeleton sequence. Target decay rate τ used in the BYOL-based learning is set to 0.99, which is verified in the following ablation study.

Linear Evaluation Protocol: The online network is frozen, and a fully connected layer (FC) is appended to online network and trained for action recognition task. Cross Entropy loss of action recognition is used as the objective.

C. Ablation Study

The proposed U-FEFP is extensively evaluated including evaluation of the proposed online network, combining the BYOL learning and pretext task-based learning, and evaluation of the overfitting under different methods. The experiments on the NTU-60 dataset are used for all the ablation studies.

Evaluation of the proposed online network: To validate the proposed spatial-temporal feature transformation network as the online network, the BYOL scheme is used with different models as online network for comparison, including MS-G3D [73], CTR-GCN [71], 2s-AGCN [19] and ST-GCN [18]. For ST-GCN, two versions are tested to further verify the overfitting problem due to a large network. The ST-GCN-v1 [18] reduces the network by reducing the number of neurons in each layer, $1/4$ to be specific, while ST-GCN-v2 reduces the network by reducing the depth of the network from 10 layers to 8 layers of ST-GCN blocks. All configurations are trained from scratch in the same way as the proposed method.

The comparison results are listed in Table I. From the results, it can be seen that both ST-GCN-v1 and ST-GCN-v2 perform better than other supervised methods [19, 71, 73], further demonstrating the overfitting problem. The proposed online network further outperforms ST-GCN-v2 [18] and achieves the best result in the BYOL-based learning. Moreover, ablation experiment on different structural compositions of our feature transformation network is also conducted. Different layers of CTR-GCN and GConv-GRU are used to construct different versions of the online network, including

TABLE II: Comparison of different modules in the proposed method

Method	X-Sub (%)	X-View (%)
Proposed online network(LSTM) + BYOL	81.21	84.50
Proposed online network(LSTM) + pretext task	62.42	67.26
Proposed online network + BYOL	84.85	88.80
Proposed online network + pretext task	70.82	75.66
U-FEFP (LSTM)	83.40	86.10
pretext task+SimCLR [74]	85.20	89.50
pretext task+MoCo [75]	86.10	90.30
U-FEFP	**86.70**	**91.20**

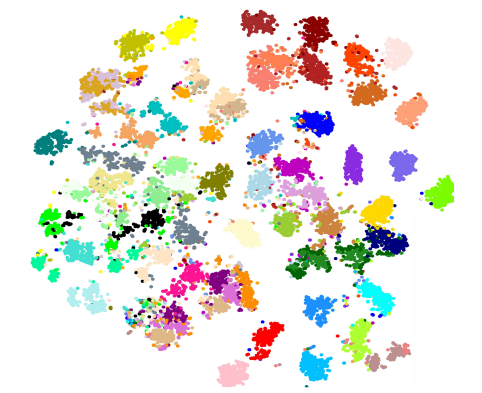


Fig. 6: t-SNE visualization of embedding for U-FEFP on the NTU-60 X-View task.

- v1: 8 CTR-GCN layers + 1 GConv-GRU layer
- v2: 6 CTR-GCN layers + 1 GConv-GRU layer
- v3: 2 CTR-GCN layers + 1 GConv-GRU layer
- v4: 5 CTR-GCN layers + 0 GConv-GRU layer (temporal pooling instead)
- v5: 5 CTR-GCN layers + 2 GConv-GRU layer
- Proposed: 5 CTR-GCN layers + 1 GConv-GRU layer

The comparison results are also shown in Table I. It can be seen that the proposed online network with 5 CTR-GCN layers + 1 GConv-GRU layer performs the best. It can also be seen that when increasing the layers of CTR-GCN or GConv-GRU over the proposed one, the performance can no longer be improved. This behaves differently to the supervised learning networks such as the CTR-GCN [71] with deep layers, indicating that it tends to be overfitting for unsupervised skeleton action recognition learning as described in Section III. Moreover, it cannot extract effective features with too few layers of CTR-GCN such as online network v3. This validates the effectiveness of our feature transformation network in extracting the spatial-temporal features and in reducing the overfitting (with less parameters than CTR-GCN [71]). Also from the perspective of computation, in practical use, only the proposed online network and the final output layer for recognition is needed and thus takes less complexity than the supervised CTR-GCN [71]. Therefore, five CTR-GCN layers and one GConv-GRU layer are used for the proposed online network in the following experiments.

Evaluation of combining BYOL-based learning and pretext task-based learning: As discussed in the Motivation, *rich distributed spatial-temporal* features *containing all information of the original skeleton* need to be generated in unsupervised learning. In order to validate this, the proposed U-FEFP is compared with the two separate modules, proposed online network with BYOL-based learning and proposed online network with pretext task-based learning. Meanwhile, to verify the effectiveness of GConv-GRU, LSTM, instead of GConv-GRU, is also used to construct the proposed method for comparison. The results are shown in Table II. The proposed U-FEFP combining the BYOL and pretext task-based learning outperforms the two separate modules, validating they can complement each other. The proposed U-FEFP using GConv-GRU outperforms U-FEFP using LSTM, verifying its effectiveness. Moreover, the results of BYOL-based learning significantly outperforms the pretext task-based learning, validating our argument in Motivation that rich distributed features in unsupervised learning matters the most since skeleton is already high-level and low-dimension features. To further verify using the BYOL in learning rich distributed features, the existing unsupervised learning approach including SimCLR [74] and Moco [75] are tested for comparison, with the pretex task based learning. The results are also shown in Table II. It can be seen that MoCo combining with pretext task is better than SimCLR, and the proposed method with BYOL outperforms both SimCLR and MoCo frameworks.

Evaluation of the overfitting under different methods: In order to illustrate the overfitting problem described in Section III, the unsupervised pre-training, fine-tuning and test processes of the proposed U-FEFP, 2s-AGCN, AS-CAL, MS2L and P&C are visualized in Fig. 1. The fine-tuning and test processes refer to the fine-tuning and test in the linear evaluation protocol where only one last FC is trained. In Fig. 1, the unsupervised pre-training process is characterized in terms of loss while the fine-tuning and test processes are in terms of accuracy and the test accuracy is achieved for each epoch in the fine-tuning process. In order for better illustration, losses of AS-CAL, MS2L, P&C are divided by 10 since the loss reduction trend is more important for comparison than the detailed value. By comparing the unsupervised pre-training and fine-tuning/test in Fig. 1, it can be seen that while 2s-AGCN achieves much smaller loss in the unsupervised pre-training stage, it performs significantly worse than the proposed U-FEFP in the fine-tuning/test in the linear evaluation protocol stage, validating its overfitting. Moreover, by comparing the fine-tuning and test in Fig. 1, it can be seen that the accuracy gaps between fine-tuning and test of 2s-AGCN, AS-CAL, MS2L and P&C are much larger than that of the proposed U-FEFP. By contrast, the performance of the proposed U-FEFP with a very small gap between fine-tuning and test is significantly better than 2s-AGCN, AS-CAL, MS2L, and P&C. This demonstrates that the proposed U-FEFP is much less prone to overfitting compared to the existing methods.

D. Comparison with the State-of-the-Art Methods

The proposed U-FEFP is compared with existing state-of-the-art methods on the different datasets. The comparison results on the NTU-60 are listed in Table III, where the proposed U-FEFP outperforms the existing unsupervised

TABLE III: Experimental results (accuracy) on the NTU-60

Method	Backbone	Train manner	X-Sub (%)	X-View (%)
Lie group [49]	-	supervised	50.10	52.80
H-RNN [76]	RNN	supervised	59.10	64.00
PA-LSTM [42]	LSTM	supervised	62.90	70.30
ST-LSTM+TS [77]	LSTM	supervised	69.20	77.70
STA-LSTM [17]	LSTM	supervised	73.40	81.20
Visualize CNN [78]	CNN	supervised	76.00	82.60
C-CNN+MTLN [10]	CNN	supervised	79.60	87.70
VA-LSTM [79]	LSTM	supervised	79.20	88.30
IndRNN [14]	RNN	supervised	81.80	88.00
ST-GCN [18]	GCN	supervised	81.50	88.30
HA-GNN [80]	GCN	supervised	93.40	97.20
LongT GAN [23]	GRU	unsupervised	39.10	48.10
PCRP [81]	GRU	unsupervised	53.90	63.50
ASCAL [27]	LSTM	unsupervised	58.50	64.80
MS^{2}L [24]	GRU	unsupervised	52.60	-
P&C [22]	GRU	unsupervised	50.70	76.30
CRRL [30]	GRU	unsupervised	67.60	73.80
SKT [36]	ST-GCN	unsupervised	72.60	77.10
ISC [29]	GRU+AGCN	unsupervised	76.30	85.20
CrossSCLR [28]	ST-GCN	unsupervised	72.90	79.90
CrossSCLR (3S) [28]	ST-GCN	unsupervised	77.80	83.40
SRCL [35]	ST-GCN	unsupervised	77.30	82.50
SRCL (3S) [35]	ST-GCN	unsupervised	80.90	85.60
ST-CL [68]	non-local GCN	unsupervised	68.10	69.40
CrossMoCo [67]	ST-GCN	unsupervised	78.40	84.90
HaLP [32]	GRU	unsupervised	79.70	86.80
3s-ActCLR [66]	ST-GCN	unsupervised	84.30	88.80
HSTM [82]	LSTM	unsupervised	82.10	91.0
3s-AimCLR++ [83]	ST-GCN	unsupervised	80.90	85.40
Skeleton2vec [84]	Transformer	unsupervised	85.70	90.30
3s-Skeleton-logoCLR [85]	ST-GCN	unsupervised	86.10	89.8
MMFR [86]	Transformer	unsupervised	84.18	89.45
3s-PSTL [39]	ST-GCN	unsupervised	79.10	82.60
3s-SkeAttnCLR [40]	ST-GCN	unsupervised	82.00	86.50
3s-HYSP [38]	ST-GCN	unsupervised	79.10	85.20
2s-ViA [87]	GCN	unsupervised	78.10	85.80
3s-RVTCLR+ [37]	ST-GCN	unsupervised	79.70	84.60
SCD-Net [72]	GCN+Transformer	unsupervised	86.60	91.70
3s-PCM [33]	GRU	unsupervised	87.40	93.10
3s-Eq-Contrast [34]	GRU	unsupervised	87.00	92.90
MAMP [65]	Transformer	unsupervised	84.90	89.10
U-FEFP	ST-GCN-v2+GC-GRU	unsupervised	85.80	90.10
U-FEFP (3S)	ST-GCN-v2+GC-GRU	unsupervised	87.20	92.40
U-FEFP	CTR-GCN+GC-GRU	unsupervised	86.70	91.20
U-FEFP (3S)	CTR-GCN+GC-GRU	unsupervised	**88.20**	**93.60**

TABLE IV: Experimental results (accuracy) on the NTU-120

Method	Backbone	Train manner	X-Sub (%)	X-Set (%)
PA-LSTM [42]	LSTM	supervised	25.50	26.30
SkeMotion [88]	LSTM	supervised	67.70	66.90
Multi CNN [89]	CNN	supervised	62.20	61.80
ST-GCN [18]	GCN	supervised	70.70	73.20
HA-GNN [80]	GCN	supervised	89.90	91.54
ASCAL [27]	LSTM	unsupervised	48.60	48.60
CRRL [30]	GRU	unsupervised	56.20	57.00
SKT [36]	ST-GCN	unsupervised	62.60	64.30
ISC [29]	GRU-AGCN	unsupervised	67.90	67.10
CrossSCLR (3S) [28]	ST-GCN	unsupervised	67.90	66.70
SRCL [35]	ST-GCN	unsupervised	67.20	67.90
SRCL (3S) [35]	ST-GCN	unsupervised	71.80	72.90
ST-CL [68]	non-local GCN	unsupervised	54.20	55.60
HaLP [32]	GRU	unsupervised	71.10	72.20
3s-ActCLR [66]	ST-GCN	unsupervised	74.30	75.70
3s-PSTL [39]	ST-GCN	unsupervised	69.20	70.30
3s-SkeAttnCLR [40]	ST-GCN	unsupervised	77.10	80.00
3s-HYSP [38]	ST-GCN	unsupervised	64.50	67.30
3s-AimCLR++ [83]	ST-GCN	unsupervised	70.10	71.20
HSTM [82]	LSTM	unsupervised	73.20	74.60
Skeleton2vec [84]	Transformer	unsupervised	79.70	81.30
3s-Skeleton-logoCLR [85]	ST-GCN	unsupervised	79.80	80.10
MMFR [86]	Transformer	unsupervised	77.09	78.82
2s-ViA [87]	GCN	unsupervised	69.20	66.90
3s-RVTCLR+ [37]	ST-GCN	unsupervised	68.00	68.90
SCD-Net [72]	GCN+Transformer	unsupervised	76.90	80.10
3s-PCM [33]	GRU	unsupervised	80.00	81.20
3s-Eq-Contrast [34]	GRU	unsupervised	79.40	81.20
MAMP [65]	Transformer	unsupervised	78.60	79.10
U-FEFP	ST-GCN-v2+GC-GRU	unsupervised	77.30	78.50
U-FEFP (3S)	ST-GCN-v2+GC-GRU	unsupervised	80.02	81.20
U-FEFP	CTR-GCN+GC-GRU	unsupervised	78.25	79.60
U-FEFP (3S)	CTR-GCN+GC-GRU	unsupervised	**80.56**	**81.67**

TABLE V: Experimental results (accuracy) on the PKU-MMD

Method	Backbone	Train manner	part I (%)	part II (%)
ST-GCN [18]	GCN	supervised	84.10	48.20
LongT GAN [23]	GRU	unsupervised	67.70	27.00
MS^{2}L [24]	GRU	unsupervised	64.90	27.60
ISC [29]	GRU+AGCN	unsupervised	80.90	36.00
CRRL [30]	GRU	unsupervised	82.10	41.80
CrossSCLR (3S) [28]	ST-GCN	unsupervised	84.90	-
SRCL [35]	ST-GCN	unsupervised	87.40	48.10
SRCL (3S) [35]	ST-GCN	unsupervised	88.20	53.20
3s-ActCLR [66]	ST-GCN	unsupervised	90.00	55.90
HaLP [32]	GRU	unsupervised	-	43.50
U-FEFP	CTR-GCN+GC-GRU	unsupervised	91.26	54.16
U-FEFP (3S)	CTR-GCN+GC-GRU	unsupervised	**92.90**	**58.90**

methods [22–24, 27–29, 33–36, 67, 68]. The proposed U-FEFP only using joint is even better than method [28] using joint, bone and motion data (3S) together. The proposed U-FEFP even performs better than some supervised learning-based methods [14, 17, 18, 42, 49, 76, 78]. Moreover, ST-GCN is also tested as backbone and it performs better than the existing unsupervised methods using ST-GCN, although slightly worse than using CTR-GCN.

Furthermore, t-SNE [70] is used to visualize the embedding clustering produced by the proposed U-FEFP on the NTU-60 X-view task using all the data. The t-SNE illustration is shown in Fig. 6. It can be seen that proposed U-FEFP can learn more discriminative latent space. Although some samples may deviate from their action class centers, they are also away from other action classes, making them easier to be discriminated. This forms the difference between the features produced by the unsupervised learning and supervised learning, and demonstrates the importance of rich distributed features where different distributed features may be used to discriminate different samples in unsupervised learning.

The result of the proposed U-FEFP compared against the existing methods on the NTU-120 dataset is shown in Table IV. The proposed U-FEFP (3S) obtains 80.56% and 81.67% on X-Sub and X-Set, respectively, and achieves the state-of-the-art performance. U-FEFP using joint, bone and motion data is better than ST-GCN using supervised way, which demonstrates the effectiveness of U-FEFP.

The result comparison on the PKU-MMD dataset is listed in Table V. It can be seen that U-FEFP (3S) achieves 92.90% and 58.90% on PKU-MMD I and PKU-MMD II, respectively. Compared with the previous best method (i.e., 3s-ActCLR [66]), the proposed method improves by 2.90% and 3.0% on PKU-MMD I and PKU-MMD II, respectively.

TABLE VI: Experimental results (accuracy) on the UAV-Human dataset

Method	Backbone	Train manner	Accuracy (%)
ST-GCN [18]	GCN	supervised	30.25
2s-AGCN [19]	GCN	supervised	34.84
HARD-Net [90]	GCN	supervised	36.97
DGNN [20]	GCN	supervised	29.90
Shift-GCN [91]	GCN	supervised	37.98
CTR-GCN(bone) [71]	GCN	supervised	41.21
CTR-GCN(joint) [71]	GCN	supervised	41.66
CTR-GCN(joint) [71]	GCN	unsupervised	21.20
CrossSCLR (3S) [28]	ST-GCN	unsupervised	35.90
SRCL [35]	ST-GCN	unsupervised	36.10
SRCL (3S) [35]	ST-GCN	unsupervised	38.40
U-FEFP	CTR-GCN+GC-GRU	unsupervised	39.22
U-FEFP (3S)	CTR-GCN+GC-GRU	unsupervised	**42.60**

The performance is significantly improved compared with the supervised ST-GCN [18], validating the effectiveness of the proposed U-FEFP.

The result comparison between the existing methods and the proposed U-FEFP on the UAV-Human dataset is shown in Table VI. It can be seen that the proposed U-FEFP achieves better performance, i.e. 42.60% using unsupervised way. From the table, it can be seen that this dataset is challenging with lower accuracy than the other used datasets. The proposed U-FEFP outperforms some methods [18, 19, 90, 91] using supervised way and unsupervised methods [28, 35], which validates the effectiveness of proposed U-FEFP in challenging scenarios.

VI. Conclusion

In this paper, we propose the U-FEFP learning framework for unsupervised skeleton-based action recognition. The U-FEFP produces rich distributed features containing all information of the skeleton sequence, which is vital for the unsupervised skeleton-based action recognition. A relatively small spatial-temporal feature transformation subnetwork combining CTR-GCN and GConv-GRU is proposed to effectively capture the skeleton sequence features. Based on this subnetwork, the unsupervised BYOL-based feature enrichment learning and unsupervised pretext task-based fidelity preservation learning are combined to formulate our U-FEFP, in order to produce the desired features. t-SNE is used to illustrate the features of the proposed U-FEFP, which demonstrates its advantage over the existing methods. Extensive experiments are conducted on NTU-60, NTU-120, PKU-MMD and UAV-Human dataset, where the proposed U-FEFP outperforms the current state-of-the-art methods and achieves the best result. Ablation study on the proposed modules is also performed and validates their effectiveness.

References

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.

[2] X.-Y. Zhang, C. Li, H. Shi, X. Zhu, P. Li, and J. Dong, "Adapnet: Adaptability decomposing encoder–decoder network for weakly supervised action recognition and localization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1852–1863, 2023.

[3] C. Pang, X. Gao, Z. Chen, and L. Lyu, "Self-adaptive graph with nonlocal attention network for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023.

[4] Y. Yang, G. Liu, and X. Gao, "Motion guided attention learning for self-supervised 3d human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8623–8634, 2022.

[5] Z. Gao, L. Guo, T. Ren, A.-A. Liu, Z.-Y. Cheng, and S. Chen, "Pairwise two-stream convnets for cross-domain action recognition with small data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 3, pp. 1147–1161, 2022.

[6] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Systems*, vol. 158, pp. 43–53, 2018.

[7] X. Gao, Y. Yang, Y. Wu, and S. Du, "Learning heterogeneous spatial–temporal context for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2023.

[8] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 445–454.

[9] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20–27.

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

[11] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, pp. 624–628, 2017.

[12] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 807–811, 2018.

[13] C. Li, Y. Hou, P. Wang, and W. Li, "Multiview-based 3-d action recognition using deep networks," *IEEE Transactions on Human-Machine Systems*, vol. 49, pp. 95–104, 2019.

[14] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5457–5466.

[15] S. Li, W. Li, C. Cook, and Y. Gao, "Deep independently recurrent neural network (indrnn)," *ArXiv*, vol.

abs/1910.06251, 2019.

[16] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 3007–3021, 2018.

[17] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Association for the Advance of Artificial Intelligence (AAAI)*, 2017, pp. 4263–4270.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Association for the Advance of Artificial Intelligence (AAAI)*, 2018, pp. 7444–7452.

[19] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 018–12 027.

[20] L. Shi, J. C. Yifan Zhang, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7904–7913.

[21] J. Kong, Y. Bian, and M. Jiang, "Mtt: Multi-scale temporal transformer for skeleton-based action recognition," *IEEE Signal Processing Letters*, vol. 29, pp. 528–532, 2022.

[22] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9628–9637.

[23] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Association for the Advance of Artificial Intelligence (AAAI)*, 2018, pp. 2644–2651.

[24] L. Lin, S. Song, W. Yang, and J. Liu, "Ms2l: Multi-task self-supervised learning for skeleton based action recognition," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 2490–2498.

[25] J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *ArXiv*, vol. abs/2006.07733, 2020.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[27] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, 2021.

[28] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4739–4748.

[29] F. M. Thoker, H. Doughty, and C. G. M. Snoek, "Skeleton-contrastive 3d action representation learning," in *ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 1655–1663.

[30] P. Wang, J. Wen, C. Si, Y. tao Qian, and L. Wang, "Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 6224–6238, 2022.

[31] M. Wang, X. Li, S. Chen, X. Zhang, L. Ma, and Y. Zhang, "Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 3207–3220, 2024.

[32] A. Shah, A. Roy, K. Shah, S. Mishra, D. Jacobs, A. Cherian, and R. Chellappa, "Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 846–18 856.

[33] J. Zhang, L. Lin, and J. Liu, "Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 7175–7183.

[34] L. Lin, J. Zhang, and J. Liu, "Mutual information driven equivariant contrastive learning for 3d action representation learning," *IEEE Transactions on Image Processing*, vol. 33, pp. 1883–1897, 2024.

[35] W. Zhang, Y. Hou, and H. Zhang, "Unsupervised skeleton-based action representation learning via relation consistency pursuit," *Neural Computing and Applications*, vol. 34, pp. 20 327–20 339, 2022.

[36] H. Zhang, Y. Hou, and W. Zhang, "Skeletal twins: Unsupervised skeleton-based action representation learning," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[37] Y. Zhu, H. Han, Z. Yu, and G. Liu, "Modeling the relative visual tempo for self-supervised skeleton-based action recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13 867–13 876.

[38] L. Franco, P. Mandica, B. Munjal, and F. Galasso, "Hyperbolic self-paced learning for self-supervised skeleton-based action representations," in *International Conference on Learning Representations*, vol. abs/2303.06242, 2023.

[39] Y. Zhou, H. Duan, A. Rao, B. Su, and J. Wang, "Self-supervised action representation learning from partial spatio-temporal skeleton sequences," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, p. 3825–3833, Jun. 2023.

[40] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part aware contrastive learning for self-supervised action recognition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 08 2023, pp. 855–863.

[41] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 3427–3431.

[42] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.

[43] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, 2020, pp. 1–24.

[44] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. yu Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2684–2701, 2020.

[45] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 261–16 270.

[46] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 3047–3060, 2020.

[47] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 081–14 097, 2023.

[48] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[49] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[50] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *International Conference on Pattern Recognition (ICPR)*, 2014, pp. 4513–4518.

[51] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2014, pp. 1–8.

[52] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Processing Letters*, vol. 25, pp. 1044–1048, 2018.

[53] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 2206–2216, 2021.

[54] R. Xia, Y. Li, and W. Luo, "Laga-net: Local-and-global attention network for skeleton based action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 2648–2661, 2022.

[55] K. Zhu, R. Wang, Q. Zhao, J. Cheng, and D. Tao, "A cuboid cnn model with an attention mechanism for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 22, pp. 2977–2989, 2020.

[56] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, and Y. Zhang, "Skeleton-based action recognition with gated convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 3247–3257, 2019.

[57] X. Jiang, K. Xu, and T. Sun, "Action recognition scheme based on skeleton representation with ds-lstm network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 2129–2140, 2020.

[58] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *IEEE Transactions on Multimedia*, vol. 20, pp. 2330–2343, 2018.

[59] W. Ng, M. Zhang, and T. Wang, "Multi-localized sensitive autoencoder-attention-lstm for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1678–1690, 2022.

[60] Y. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 1915–1925, 2020.

[61] C. Wu, X. Wu, and J. Kittler, "Graph2net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 2120–2132, 2021.

[62] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan, "A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 64–76, 2021.

[63] Y. B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, and L. Lin, "Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[64] B. Kim, H. J. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 209–225.

[65] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li, "Masked motion predictors are strong 3d action representation learners," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 10 181–10 191.

[66] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action

recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2363–2372.

[67] Q. Zeng, C. Liu, M. Liu, and Q. Chen, "Contrastive 3d human skeleton action representation learning via crossmoco with spatiotemporal occlusion mask data augmentation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1564–1574, 2023.

[68] X. Gao, Y. Yang, Y. Zhang, M. Li, J.-G. Yu, and S. Du, "Efficient spatio-temporal contrastive learning for skeleton-based 3-d action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 405–417, 2023.

[69] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7559–7576, 2023.

[70] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[71] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 13 339–13 348.

[72] C. Wu, X. Wu, J. Kittler, T. Xu, S. Atito, M. Awais, and Z. Feng, "Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2024, pp. 5949–5959.

[73] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 140–149.

[74] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[75] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[76] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[77] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 3007–3021, 2017.

[78] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.

[79] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2136–2145.

[80] P. Geng, X. Lu, W. Li, and L. Lyu, "Hierarchical aggregated graph neural network for skeleton-based action recognition," *IEEE Transactions on Multimedia*, pp. 1–16, 2024.

[81] S. Xu, H. Rao, X. Hu, J. Cheng, and B. Hu, "Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 624–634, 2020.

[82] W. Cao, A. Zhang, Z. He, Y. Zhang, and X. Yin, "Hierarchical spatial-temporal masked contrast for skeleton action recognition," *IEEE Transactions on Artificial Intelligence*, pp. 1–14, 2024.

[83] T. Guo, M. Liu, H. Liu, G. Wang, and W. Li, "Improving self-supervised action recognition from extremely augmented skeleton sequences," *Pattern Recognition*, vol. 150, p. 110333, 2024.

[84] R. Xu, L. Huang, M. Wang, J. Hu, and W. Deng, "Skeleton2vec: A self-supervised learning framework with contextualized target representations for skeleton sequence," 2024.

[85] J. Hu, Y. Hou, Z. Guo, and J. Gao, "Global and local contrastive learning for self-supervised skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–12, 2024.

[86] X. Zhu, X. Shu, and J. Tang, "Motion-aware mask feature reconstruction for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–12, 2024.

[87] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Brémond, "View-invariant skeleton action representation learning via motion retargeting," *International Journal of Computer Vision*, p. 2351–2366, 2024.

[88] J. Liu, G. Wang, P. Hu, L. yu Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3671–3680.

[89] Q. Ke, Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, vol. 27, pp. 2842–2855, 2018.

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

[91] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 180–189.

Chuankun Li received the Ph.D degree with School of electronic information engineering , Tianjin University, China in 2020. He is currently an associate professor with the School of Information and Communication Engineering for North University of China. His current research interests include computer vision and machine learning.

Shuai Li is currently with the School of Control Science and Engineering, ShanDong University (SDU), China, as a Professor and QiLu Young Scholar. He was with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, China, as an Associate Professor from 2018-2020. He received his Ph.D. degree from the University of Wollongong, Australia, in 2018. His research interests include image/video coding, 3D video processing and computer vision. He was a co-recipient of two best paper awards at the IEEE BMSB 2014 and IIH-MSP 2013, respectively.

Yanbo Gao is currently with the School of Software, Shandong University (SDU), Jinan, China, as an Associate Professor. She was with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, as a Post-doctor from 2018-2020. She received her Ph.D. degree from UESTC in 2018. Her research interests include video coding, 3D video processing and light field image coding. She was a co-recipient of the best student paper awards at the IEEE BMSB 2018.

Xingyu Gao (Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has been a Visiting Scholar with Nanyang Technological University, Singapore, and Singapore Management University, Singapore. He is currently a Professor with the Institute of Microelectronics, Chinese Academy of Sciences. His current research interests include machine learning, multimedia content analysis and retrieval, and computer vision.

Ping Chen received the Ph.D. degree in signal and information processing from the North University of China, Taiyuan, China in 2011. He is currently a professor with the School of Information and Communication Engineering for North University of China. His research interests include X-ray CT imaging and CT reconstruction algorithm, image processing and recognition.

Jian Li received the Ph.D. degree in signal and information processing from the North University of China, Taiyuan, China. He is currently a professor with the School of Information and Communication Engineering for North University of China.His research interests include acquisition and processing of arrayed signals such as vibration, sound, and shockwave, image processing and reconstruction, and development of embedded systems.

Wanqing Li (M'97-SM'05) received his Ph.D. in electronic engineering from The University of Western Australia. He was a Principal Researcher (98-03) at Motorola Research Lab and a visiting researcher (08, 10 and 13) at Microsoft Research US. He is currently an Associate Professor and Co-Director of Advanced Multimedia Research Lab (AMRL) of UOW, Australia. His research areas are machine learning, 3D computer vision, 3D multimedia signal processing and medical image analysis. Dr. Li currently serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON MULTIMEDIA. He was an Associator for JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION.