Momentum Contrastive Teacher for Semi-Supervised Skeleton Action Recognition

Mingqi Lu, Xiaobo Lu and Jun Liu

***Abstract*—In the field of semi-supervised skeleton action recognition, existing work primarily follows the paradigm of self-supervised training followed by supervised fine-tuning. However, self-supervised learning focuses on exploring data representation rather than label classification. Inspired by MeanTeacher, we explore a novel pseudo-label-based model called SkeleMoCLR. Specifically, we use MoCo v2 as the foundation and extend it into a teacher-student network through a momentum encoder. The generation of high-confidence pseudo-labels requires a well-pretrained model as a prerequisite. In cases where large-scale skeleton data is lacking, we propose leveraging contrastive learning to transfer discriminative action features from large vision-text models to the skeleton encoder. Following the contrastive pre-training, the key encoder branch from MoCo v2 serves as the teacher to generate pseudo-labels for training the query encoder branch. Furthermore, we introduce pseudo-labels into the memory queues, sampling negative samples from different pseudo-label classes to maximize the representation differentiation between different categories. We jointly optimize the classification loss for both labeled and pseudo-labeled data and the contrastive loss for unlabeled data to update model parameters, fully harnessing the potential of pseudo-label semi-supervised learning and self-supervised learning. Extensive experiments conducted on the NTU-60, NTU-120, PKU-MMD, and NW-UCLA datasets demonstrate that our SkeleMoCLR outperforms existing competitive methods in the semi-supervised skeleton action recognition task.**

***Index Terms*—Action recognition, Skeleton, Semi-supervised, Contrastive learning.**

I. Introduction

Skeletal data has witnessed a surge in popularity for the analysis of human activities in recent years. Unlike RGB frames and depth maps, skeletal data stands out for its lightweight nature and remarkable robustness to changes in lighting, texture, and background conditions. Supervised skeleton action recognition approaches heavily rely on a large amount of labeled data, which is costly and labor-intensive to collect.

This work was supported in part by the National Natural Science Foundation of China under Grants 62271143, in part by the Big Data Computing Center of Southeast University. (Corresponding author: Xiaobo Lu.)

Mingqi Lu is with the School of Automation, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China, and also with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore (e-mail: lumingqiseu@gmail.com).

Xiaobo Lu is with the School of Automation, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: xblu2013@126.com).

Jun Liu is with School of Computing and Communications, Lancaster University, UK (e-mail: j.liu81@lancaster.ac.uk).

Semi-supervised learning aims to leverage a small amount of labeled data and a large volume of unlabeled data. It holds significant value in practical domains where acquiring labeled data is costly. Recent developments in semi-supervised learning can be primarily categorized into two main types. The first type is pseudo-labeling, also known as self-training. In pseudo-labeling, the model generates class predictions for each unlabeled sample, and these predictions are used as artificial labels for training, such as FixMatch [1]. The second type involves utilizing self-supervised learning on the unlabeled data, followed by supervised fine-tuning on the labeled data, such as SimCLR [2] and MoCo [3]. However, self-supervised learning aims to explore the inherent representations of data itself instead of label prediction. In the field of semi-supervised skeletal action recognition, representative methods use MoCo v2 [4] framework to extract discriminative features from pre-augmented skeleton actions, such as CrossCLR [5] and AimCLR [6]. However, they heavily rely on strong skeletal data augmentation, which may compromise the inherent structural information of skeletal actions, leading to information bias. Differing from previous methods, we attempt to address the semi-supervised skeleton action recognition task from the perspective of pseudo-labeling. However, without well-established pre-trained models providing implicit constraints, semi-supervised learning based on pseudo-labeling can easily be misled by inaccurate pseudo-labels, especially in cases with a large label space. Compared to the network-scale visual-text data, widely used skeleton datasets are relatively much smaller. Acquiring a well-pretrained skeletal encoder presents a challenging task.

Skeleton sequences are typically present alongside RGB videos, whether they originate from pose estimation algorithms or depth cameras like Kinect. Skeletal information is concise and robust; however, the lack of detailed body information (such as appearance and objects) can lead to difficulties in handling similar and complex actions. Classic multimodal methods model the skeletal data and RGB images separately using two networks to compensate for the shortcomings of a single modality, but they require substantial computational resources during both training and inference stages. However, due to modality differences, it is not directly possible to establish consistency constraints between skeletons and videos. Large vision-text pre-training models like CLIP [7] achieve significant success in the visual domain. ActionCLIP [8] is a classic derivative model of CLIP in the field of video action recognition. In our pre-training method, we employ contrastive learning to facilitate the feature transfer from large-scale pre-trained vision models to the skeleton

encoder, thereby enabling the learning of more comprehensive and extensive representations. Specifically, we treat one-to-one corresponding skeletons and videos as positive samples, and the frozen encoder in ActionCLIP extracts video features. We utilize a contrastive loss in a low-dimensional space to align skeleton embeddings and video embeddings, thereby training the skeleton encoder. In our semi-supervised model, we load the weights from the contrastive pre-training as the initialization for the skeleton encoder. Unlike methods that use multimodal data during both the training and inference stages, we introduce video-text data only during the contrastive pre-training phase. In the subsequent stages, only skeleton data is required as input for action recognition, without the involvement of visual and language modalities, thus incurring no additional computational costs.

Mean Teacher [9] generates target samples through exponential moving averages. Inspired by this, we propose SkeleMoCLR for semi-supervised skeleton action recognition, based on MoCo V2, as illustrated in Figure 1. For each set of skeleton inputs, the key encoder, along with an additional classifier, serves as the teacher network, generating pseudo-labels required for training the query encoder. Simultaneously, the projectors map positive and negative representations onto a low-dimensional embedding, facilitating discriminative learning.Since the selection of negative samples is crucial in contrastive learning, we incorporate class information from pseudo-labels into the memory queue of MoCo V2. For a query sample, samples with different pseudo-labels constitute negative examples. This is to maximize the distinctiveness in the representations between query samples and negative samples from different classes, thus achieving category contrast based on pseudo-labels. Furthermore, we introduce a regularization term between labeled and unlabeled representations to enhance their consistency and the model's generalization ability. We jointly optimize the cross-entropy loss for labeled data and pseudo-labeled data, as well as the contrastive loss for unlabeled data, updating model parameters using gradients from both labeled and unlabeled data.

Our main contributions can be summarized as follows:

- This paper neatly extends MoCo v2 into a teacher-student network using momentum updates, and proposes a novel semi-supervised skeleton action recognition model, SkeleMoCLR.
- This paper uses large-scale vision-text models as a bridge and introduces a contrastive pretraining strategy to initialize the weights of the skeleton encoder used for generating pseudo-labels.
- This paper incorporates class information from pseudo-labels into contrastive learning, and facilities the learning of more discriminative class-level representations through negative sample sampling.
- SkeleMoCLR achieves performance comparable to state-of-the-art methods on the semi-supervised skeletal action recognition task on NTU-60, NTU-120, PKU-MMD and NW-UCLA datasets.

II. Related Work

A. Human Action Recognition

Human action recognition is vital for video understanding. Both skeletal sequences and RGB videos are widely used input modalities for human action recognition. While skeleton information is concise and robust, attracting significant attention to skeleton action recognition, the lack of detailed body information limits performance. Various modalities have been explored to address these limitations, including RGB images [10], textual descriptions [11], and depth images [12]. Classic multimodal approaches integrate the prediction results of skeleton data and other modalities (e.g., RGB images and depth images), requiring substantial computational resources and exhibiting inefficiency in both training and inference stages. PoseConv3D [13] uses both RGB heatmaps and skeletal modalities for robust human action recognition. VPN [14] projects 3D poses and their corresponding RGB videos into a common embedding space, learning spatiotemporal relationships through an attention network. In the absence of large-scale skeletal data, we propose leveraging large vision-text models for pre-training the skeleton encoder to enable efficient inference with a single skeleton modality, generating informative features based on the strong generalization capabilities of the vision-text model.

B. Supervised Skeleton Action Recognition

RNN-based models [15], [16] treat skeleton data as extensive sequential data. CNN-based techniques [17], [18] transmute skeleton sequences into image-like representations. The introduction of GCN has paved the way for innovations like ST-GCN [19], which mold skeletal data into predefined spatial graphs and employ GCNs to amalgamate joint information. A multitude of GCN-based approaches [20]–[22] have sprung up, further building upon the foundation laid by ST-GCN. ShiftGCN++ [23] introduces lightweight spatial and temporal shift graph convolutions. FGCN [24] incorporates a feedback mechanism into GCN for action recognition. 2s-AGCN [22] adopts a dual-stream approach for joint and bone and introduces adaptive dynamic learning module. Due to its strong and robust spatiotemporal feature extraction capabilities, we employ 2s-AGCN as the skeleton encoder in SkeleMoCLR.

C. Self-Supervised Skeleton Action Recognition

Self-supervised learning aims to learn discriminative representations from a large amount of unlabeled data. LongT GAN [25] utilizes a cyclic encoder-decoder GAN for reconstructing input sequences. In a similar vein, Predict&Cluster [26] introduces a decoder to enhance representation capabilities. MS2L [27] presents a multi-task self-supervised learning framework that incorporates motion prediction. ISC [28] introduces an approach that combines sequence-based and graph-based skeleton contrastive methods. CRRL [29] utilizes contrastive reconstruction to capture both pose and motion features. CrosSCLR [5] employs a cross-view knowledge mining strategy to capture more comprehensive representations. AimCLR [6] focuses on extensive action augmentation to compel the encoder to

learn more general representations. However, skeletal data is very concise, and tasks such as reconstruction and strong transformations can harm the topological information of the skeleton, resulting in biases.

D. Semi-Supervised Skeleton Action Recognition

Currently, there is relatively limited research in the field of semi-supervised skeleton action recognition. ASSL [30] employs adversarial regularization to align the features of both labeled and unlabeled data. CD-JBF-GCN [31] integrates both joint and bone information by facilitating the transmission of motion details. Xu et.al [32] uses contrastive learning to extract and aggregate action representations at the body, part, and joint levels. MAC-Learning [33] uses anchor graphs to create soft positive/negative pairs and proposes multi-granularity contrastive losses. To capture more semantic information, X-CAR [34] explores consistent action representations between joint data and learnable augmented data through contrastive learning. Current methods primarily rely on skeletal transformations to generate positive and negative pairs separately for exploring action representations through contrastive learning. However, it is surprising that mainstream pseudo-labeling methods in semi-supervised learning have not yet been reported in the literature. Our work aims to fill this gap.

III. Methodology

Figure 1 depicts the network architecture of SkeleMoCLR, designed for semi-supervised skeleton action recognition. SkeleMoCLR uses MoCo v2 as a foundation and extends it into a teacher-student network through a momentum encoder. We utilize a large vision-text model for contrastive pre-training of the skeletal encoder. After pre-training, the key encoder branch of MoCo v2 serves as the teacher to generate pseudo-labels for training the query encoder branch. Additionally, we introduce pseudo-labels into the memory queue and sample negative examples from different pseudo-label classes to maximize representation differences among various categories. Subsequent sections provide separate explanations of the contrastive pre-training, semi-supervised framework, and representation regularization components within our approach.

A. Contrastive Pre-training

As shown in Figure 2, our contrastive pretraining framework is based on MoCo v2. It creates positive pairs (x_s, x_v) from a skeleton sequence x_s and its corresponding RGB video x_v, and then generates embeddings (ζ_s, ζ_v) separately through an encoder and a projector. The encoders in ActionCLIP remain frozen, with only the parameters of the skeleton encoder being trainable. We maintain a memory queue $\tilde{Z}$ to store negative samples, and the cross-modal contrastive loss is

$$L(s,v) = -\log \frac{\exp\left(\zeta_s \cdot \zeta_v/\tau\right)}{\exp\left(\zeta_s \cdot \zeta_v/\tau\right) + \sum_{i=1}^{N} \exp\left(\zeta_s \cdot \tilde{\zeta}_v^i/\tau\right)} \tag{1}$$

Where $\zeta_s \cdot \zeta_v$ represents the normalized dot product. $\tilde{\zeta}_v^i$ represents the embedding in $\tilde{Z}$ corresponding to the i-th negative sample. N represents the quantity of negative features, and τ is a temperature hyperparameter. The parameters of the skeleton projector and video projector are denoted as θ_s and θ_v, updated using the following equation: $\theta_v \leftarrow \alpha\theta_v + (1-\alpha)\theta_s$, where $\alpha \in [0, 1)$ represents the momentum coefficient. MLP projectors are completely discarded after pre-training. In the subsequent tasks, the skeletal encoder initializes its weights by loading the weights from the contrastive pretraining.

B. Semi-Supervised Framework

Our semi-supervised skeleton action recognition model is illustrated in Figure 1. We consider the encoders with momentum updates from MoCo v2 as a teacher-student network for generating pseudo-labels. Given a skeletal sequence s, we apply data augmentation to construct positive pairs (s, s'). Following MoCo v2, we train two skeletal encoders: a query encoder f_q and a key encoder f_k. For each pair (s, s'), skeletal embeddings (ζ_q, ζ_k) are generated through the skeletal encoder and projector. As the parameters of the key encoder are a momentum-updated version of the query encoder, $f_k \leftarrow \alpha f_k + (1-\alpha) f_q$, the key encoder produces more stable representations throughout the entire training, improving the optimization process. Therefore, we employ the key network as the teacher and the query network as the student in our semi-supervised framework. During each iteration, the teacher network generates pseudo-labels $\hat{p}^u$ for normally augmented data. The student network calculates prediction probabilities p^u for extremely augmented versions of the same skeletal sequence, and use the high-confidence pseudo-labels $\hat{p}^u$ as the targets in the cross-entropy loss for unlabeled data.

In standard contrastive learning, negative samples are samples other than the positive samples. We consider samples from other pseudo-label classes as negative samples, thus leveraging valuable discriminative information from the pseudo-labels.

$$L_{CL} = -\log \frac{\exp\left(\zeta_q \cdot \zeta_k^{\hat{p}^u}/\tau\right)}{\exp\left(\zeta_q \cdot \zeta_k^{\hat{p}^u}/\tau\right) + \sum_{c=1}^{\{1,2,\cdots,C\}\setminus\hat{p}^u} \sum_{j=1}^{D} \exp\left(\zeta_q \cdot \zeta_{kj}^{c}/\tau\right)} \tag{2}$$

Where C represents the number of classes, and $\zeta_q \cdot \zeta_k$ denotes the dot product used to calculate the similarity between two normalized embeddings. ζ_{kj}^c represents negative samples stored in C queues, where the size of each queue is D. During each iteration, for unlabeled samples, the earliest samples in the corresponding queues are gradually replaced based on their pseudolabels.

As shown in Figure 1, a mini-batch is sampled consisting of labeled skeletal data and unlabeled skeletal data. Forward propagation is computed for both labeled and unlabeled skeletal sequences, and the relevant predictions and pseudo-label targets are concatenated. For the labeled skeletal data:

$$L_{CE}^{l} = -\sum_{i=1}^{C} y_i^l \log\left(p_i^l\right) \tag{3}$$

For the unlabeled skeletal data:

$$L_{CE}^{u} = -\sum_{i=1}^{C} \hat{p}_i^u \log\left(p_i^u\right) \tag{4}$$

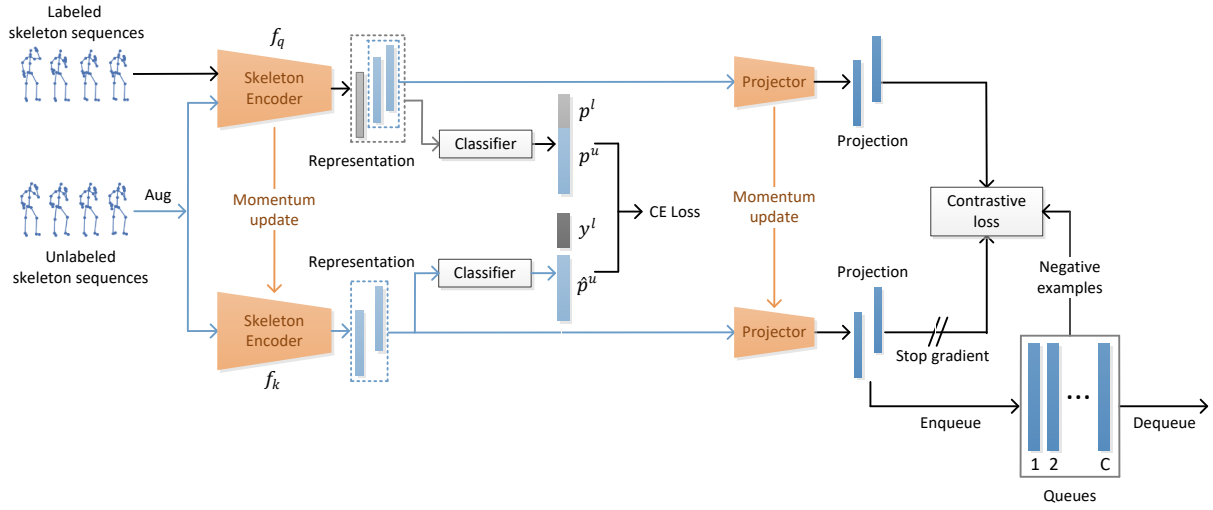


Fig. 1. The framework of SkeleMoCLR. Skeletal sequences are fed into the skeleton encoders, generating feature vectors for classification and feature embeddings for self-supervised learning, respectively. A teacher-student network is established using a momentum encoding mechanism to create pseudo-labels, and model parameters are updated through a combination of semi-supervised and contrastive losses.

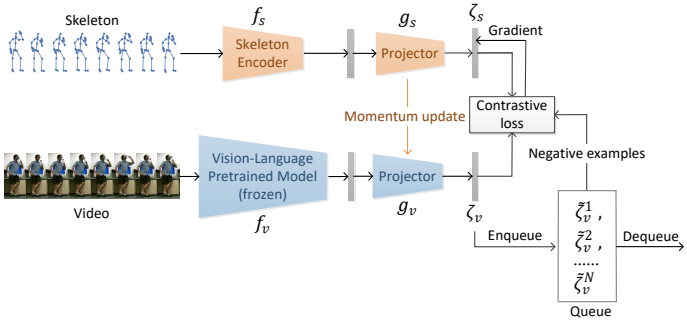


Fig. 2. An overview of the proposed framework for contrastive pre-training.

This gradual transfer of category information from the labeled dataset to the unlabeled dataset enhances the model's representation learning.

C. Representation Regularization

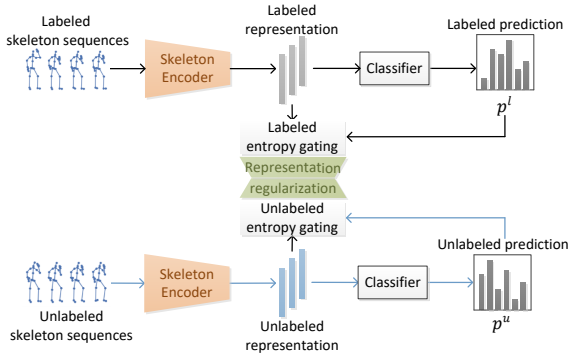


Fig. 3. Illustration of the representation regularization. This regularization term is introduced between representations of labeled and unlabeled data to enhance generalization.

Inspired by [35], we introduce a regularization term to bring the representation distribution of unlabeled and labeled samples as close as possible, as shown in Figure 3. With limited labeled data, as unlabeled samples inherently contain latent data information, we leverage these unlabeled samples to guide the training of labeled samples and learn representations with stronger generalization ability. We adopt an adaptive strategy utilizing an entropy-gating function. Only samples with high confidence (both labelled and unlabeled) are subject to the Maximum Mean Discrepancy (MMD) constraint L_R. During each iteration, high-confidence labeled and unlabeled samples within the mini-batch are added to their respective buffers. Subsequently, the latest k samples are drawn from these buffers to form the feature representation sets (H^l, H^u).

$$L_R = MMD(H^l, H^u) \tag{5}$$

In the total loss function,

$$L = L^l_{CE} + \lambda_{CE} L^u_{CE} + \lambda_{CL} L^u_{CL} + \lambda_R L_R \tag{6}$$

IV. Experiments

A. Datasets

To assess the effectiveness of the proposed method, extensive experiments are conducted on four widely used skeleton action recognition datasets: NTU-RGB+D 60 [36], NTU-RGB+D 120 [37], PKU-MMD [38], and Northwestern-UCLA [39].

NTU RGB+D. The NTU RGB+D (NTU-60) dataset [36] is a comprehensive dataset consisting of 56,578 skeletal sequences across 60 distinct action categories. These sequences are captured from 40 volunteers, each contributing data from 25 joints. The data is collected using three Microsoft Kinect v2 cameras. The evaluation is conducted following two standard evaluation protocols: cross-subject (CS) and cross-view (CV) protocols. Under the CS protocol, the training set comprises 40,091 skeleton sequences from 20 volunteers, while the test set includes 16,487 sequences from another group of 20 volunteers. In the CV protocol, the training set consists of 37,646 skeleton sequences captured by cameras 2 and 3,

and the test data encompasses 18,932 sequences recorded by camera 1. In the semi-supervised training setup, we utilize only 1%, 5%, 10%, and 20% of labeled data along with the corresponding remaining unlabeled data.

NTU RGB+D 120. The NTU RGB+D 120 (NTU-120) [37] dataset is an extension of the NTU RGB+D dataset. It comprises 113,945 samples from 120 action categories performed by 106 subjects. This dataset is well-defined through two protocols, namely Cross-Subject (CS) and Cross-Setup (CE). Under the CS protocol, data from 53 distinct subjects were utilized to assemble 63,026 training samples and 50,919 testing samples. Within the CE protocol, there are 32 different setup IDs, with 54,468/59,477 action sequences possessing even/odd setup IDs used for training/testing purposes.

PKU-MMD. PKU-MMD [38] is a multi-modal dataset designed for 3D human behavior understanding, comprising 51 different actions and a total of 21,544 video segments. The dataset is divided into two subsets: the first part includes 21,539 instances, representing a relatively simplified version, while the second part consists of 6,904 instances with more challenging variations in viewpoint. Each sample is composed of 25 body joints. We conduct experiments on these two subsets using a cross-subject protocol. In the semi-supervised setting of PKU-MMD, the training data for each category typically includes around 1% to 10% labeled data.

Northwestern-UCLA. The Northwestern-UCLA (NW-UCLA) dataset [39] comprises 1494 samples from 10 distinct action categories, captured using 3 Kinect cameras. The data collection involves 10 volunteers, each with 20 skeletal joints. The training set includes 1018 samples from the first two views, while the testing set encompasses 476 samples from the third view. For the semi-supervised scenarios, we employ only 5%, 15%, 30%, and 40% of labeled data, along with the corresponding remaining unlabeled data.

B. Experimental Settings

We employ the same data preprocessing as SkeletonCLR [5] and AimCLR [6]. Through linear interpolation, we adjust the temporal length of all skeleton sequences to a fixed length of 50 frames. In contrastive pre-training, the vision encoder is ViT-B/16, and we use the text encoder from CLIP, inputting action category names as text prompts. We use YOLOv8 as a human detector in video frames while scaling the extracted human frames to save computational and storage resources. The vision encoder samples 8 frames from the human frames to construct visual mappings. Training is conducted following the ActionCLIP methodology. In the semi-supervised setting, we adopt a category balancing strategy consistent with the majority of methods to sample labeled data. The training set also include the corresponding remaining unlabeled skeleton sequences. We adopt 2S-AGCN as the skeletal encoder. The projectors within the model are all 2-layer perceptron (MLP) with ReLU activation function, ultimately projecting features into a 128-dimensional space. The size of the memory queue and hyperparameter τ are set to 8192 and 0.07, respectively. Output vectors are normalized using L2-norm. In the semi-supervised training, the hyperparameter D is set to 64 and the threshold is 0.5. We employ two sets of augmentations: normal augmentations and extreme augmentations. The specific augmentation methods used are consistent with those outlined in [6]. We use SGD as the optimizer with a momentum of 0.9 and weight decay of 0.0004. The entire training consists of 200 epochs, with the first 10 steps following a warm-up strategy. The learning rate is scheduled to decay with cosine annealing, ranging from a maximum learning rate of 0.1 to a minimum learning rate of 0.0001. For both pretraining and downstream tasks, the batch size for NTU-60 and NTU-120 is set to 128, while for PKU-MMD and NW-UCLA, it is set to 64. All experiments are conducted on a single GeForce GTX 3090 GPU using the PyTorch framework [40].All the results reported in our experiments are from the teacher model, which employs the EMA strategy, rather than the backpropagation strategy, to update parameters. This approach significantly reduces computational requirements and training time.

C. Semi-Supervised Evaluation

In the NTU-60, PKU-MMMD, and NW-UCLA datasets, we conduct comparisons between the proposed model and current state-of-the-art methods, as shown in Tables 1, 2, and 3, respectively. The compared methods include semi-supervised based methods [41], [42], [43], [44], [30], [33], [34], and self-supervised based methods [25], [27], [28], [31], [45], [46], [47], [48], [49], [50], [5], [6], [51], [52], [53], [54], [55]. SkeleMoCLR demonstrate competitive performance across all evaluation protocols on these four datasets. During the experimental inferencing phase, we use single skeletal modality data as input for all methods. The prefixes '2s-' and '3s-' indicate models based on dual-stream (Joint+Bone) and triple-stream (Joint+Bone+Velocity) architectures, respectively. Models without a prefix use only the joint stream. "w/o PT" indicates models that are not loaded with contrastive pretraining weights, i.e., they are initialized randomly. Table 1 presents a performance comparison between relevant methods and 2s-SkeleMoCLR on the NTU-60 dataset. At 1% labeled data, 2s-SkeleMoCLR performs comparably to the state-of-the-art method 3s-SkeAttnCLR and outperforms other approaches. When using 10% labeled data, 2s-SkeleMoCLR exhibits a 0.2% performance gain over 3s-AimCLR on both the CS and CV protocols. With 20% and 40% labeled data, 2s-SkeleMoCLR's performance on the CS and CV protocols reaches state-of-the-art levels. Experimental results demonstrate that 2s-SkeleMoCLR effectively learns rich skeletal motion representations and a higher-quality feature space. Compared to random initialization, loading contrastive pre-training weights into 2s-SkeleMoCLR significantly improves semi-supervised performance, highlighting the effectiveness of this strategy.

As shown in Table 2, on the PKU-MMD dataset with only 1% labeled data, 2s-SkeleMoCLR's performance is on par with the state-of-the-art method. With 10% labeled data, 2s-SkeleMoCLR outperforms the current state-of-the-art method (3s-SkeleMixCLR) by 0.6% on PKU-MMD Part-I and by 0.3% on PKU-MMD Part-II. The second part of the PKU-MMD dataset is more challenging due to viewpoint variations,

TABLE I
RECOGNITION ACCURACY (%) ACHIEVED BY DIFFERENT METHODS ON NTU-60 WITH 1%, 5%, 10%, 20% AND 40% OF LABELED TRAINING DATA.

Method	1%		5%		10%		20%		40%	
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
S^4L [41]	-	-	48.4	55.1	58.1	63.6	63.1	71.1	68.2	76.9
Pseudolabels [42]	-	-	50.9	56.3	58.4	65.8	63.9	71.2	69.5	77.7
VAT [43]	-	-	51.3	57.9	60.3	66.3	65.6	72.6	70.4	78.6
VAT+EntMin [44]	-	-	51.7	58.3	61.4	67.5	65.9	73.3	70.8	78.9
ASSL [30]	-	-	57.3	63.6	64.3	69.8	68.0	74.7	72.3	80.0
ISC [28]	35.7	38.1	59.6	65.7	65.9	72.5	70.8	78.2	-	-
EnGAN-PoseRNN [56]	-	-	-	-	-	-	-	-	78.7	86.5
2s-CD-JBF-GCN [31]	-	-	61.8	65.3	71.7	78.0	78.4	85.9	83.2	90.9
2s-MAC-Learning [33]	-	-	63.3	70.4	74.2	78.5	78.4	84.6	81.1	89.6
X-CAR [34]	-	-	67.3	70.0	76.1	78.2	79.4	85.7	84.1	90.4
CMD [45]	50.6	53.0	71.0	75.3	75.4	80.2	78.7	84.3	-	-
HaLP [46]	46.6	48.7	66.9	71.5	72.6	77.1	76.1	82.4	-	-
HaLP+CMD [46]	52.6	53.0	71.4	75.3	76.0	80.4	79.2	84.6	-	-
SDS-CL [47]	-	-	71.3	75.3	77.2	83.0	**82.2**	86.4	86.4	91.1
HiCLR [48]	58.5	58.3	-	-	79.6	84.0	-	-	-	-
3s-Hi-TRS [49]	49.3	51.5	71.5	74.8	77.7	81.1	-	-	-	-
3s-Colorization [50]	48.3	52.5	65.7	70.3	71.7	78.9	76.4	82.7	79.8	86.8
3s-CrosSCLR [5]	51.1	50.0	-	-	74.4	77.8	-	-	-	-
3s-AimCLR [6]	54.8	54.3	-	-	78.2	81.6	-	-	-	-
3s-CMD [45]	55.6	55.5	**74.3**	**77.2**	79.0	82.4	81.8	86.6	-	-
3s-SkeleMixCLR [51]	55.3	55.7	-	-	79.9	83.6	-	-	-	-
3s-SkeAttnCLR [52]	**59.6**	59.2	-	-	81.5	83.8	-	-	-	-
2s-DMMG [53]	56.1	56.6	-	-	**81.8**	**85.1**	-	-	-	-
2s-SkeleMoCLR w/o PT	51.6	53.1	66.7	69.9	72.6	76.1	76.9	81.3	80.6	88.2
2s-SkeleMoCLR	58.5	**59.5**	72.9	75.9	78.4	81.8	**82.2**	**86.8**	**86.5**	**94.8**

and our method performs well, demonstrating strong robustness. As shown in Table 3, on NW-UCLA, 2s-SkeleMoCLR outperforms other competing methods with only 5% labeled data. Compared to the state-of-the-art method X-CAR, 2s-SkeleMoCLR shows improvements of 0.2% and 0.4% at 30% and 40% labeled data, respectively.

Our performance gains are largely attributed to the contrastive pre-training with large-scale vision-text models. However, on the smaller PKU-MMD and NW-UCLA datasets, the advantage of contrastive pre-training for the skeleton encoder is not as evident with 1% and 5% labeled data. Nonetheless, as shown in Table 1, the 2s-SkeleMoCLR without pre-trained weights still performs comparably to 3s-Colorization and 3s-CrosSCLR on the NTU-60 dataset, demonstrating the effectiveness of our semi-supervised framework design.

TABLE II
RECOGNITION ACCURACY (%) ACHIEVED BY DIFFERENT METHODS ON PKU-MMD WITH 1% AND 10% OF LABELED TRAINING DATA.

Method	1%		10%	
	Part-I	Part-II	Part-I	Part-II
LongT GAN [25]	35.8	12.4	69.5	25.7
MS^2L [27]	36.4	13.0	70.3	26.1
ISC [28]	37.7	-	72.1	-
ACL [54]	58.7	**17.7**	86.7	37.2
3s-CrosSCLR [5]	49.7	10.2	82.9	28.6
3s-AimCLR [6]	57.5	15.1	86.1	33.4
3s-SkeleMixCLR [51]	62.2	15.7	87.7	41.0
3s-PSTL [55]	**62.5**	16.9	86.9	42.0
2s-SkeleMoCLR	61.6	17.1	**88.3**	**42.3**

D. Wide-Ranging Experiments

To further validate the performance of the proposed method in self-supervised learning, a model that uses only unlabeled skeletal data as input, denoted as 2s-SkeleMoCLR-, is employed. The feature quality learned by the skeletal encoder is evaluated under four evaluation protocols and compared with state-of-the-art techniques.

TABLE III
RECOGNITION ACCURACY (%) BY DIFFERENT METHODS ON NW-UCLA WITH LABELED TRAINING DATA AT 5%, 15%, 30%, AND 40%.

Method	5%	15%	30%	40%
S^4L [41]	35.3	46.6	54.5	60.6
Pseudolabels [42]	35.6	48.9	60.6	65.7
VAT [43]	44.8	63.8	73.7	73.9
VAT+EntMin [44]	46.8	66.2	75.4	75.6
ASSL [30]	52.6	74.8	78.0	78.4
MAC-Learning [33]	63.0	**78.8**	79.9	81.6
X-CAR [34]	**68.7**	77.5	80.9	83.1
SDS-CL [47]	67.0	78.2	79.3	82.8
2s-SkeleMoCLR	68.3	78.2	**81.1**	**83.5**

KNN Evaluation. A k-nearest neighbors (KNN) classifier with no trainable parameters is used. Using only joint stream, SkeleMoCLR- is compared to other relevant methods on NTU-60, NTU-120, and PKU-MMD, as shown in Table 4 and Table 5. Our SkeleMixCLR- significantly outperforms competing methods, especially on the larger NTU-120. The substantial gains when using a parameterless classifier indicate that our method learns features with stronger discriminative power.

TABLE IV
KNN EVALUATION RESULTS FOR DIFFERENT METHODS ON NTU-60 AND NTU-120.

Method	NTU-60		NTU-120	
	CS	CV	CS	CE
LongT GAN [25]	39.1	48.1	31.5	35.5
P&C [26]	50.7	76.3	39.5	41.8
ISC [28]	62.5	82.6	50.6	52.3
CMAL [54]	64.2	72.3	50.0	52.1
SkeletonCLR [5]	64.8	60.7	41.9	42.9
CrosSCLR-B [5]	66.1	81.3	52.5	54.9
AimCLR [6]	71.0	63.7	48.9	47.3
SkeleMixCLR [51]	72.3	65.5	49.3	48.3
SkeAttnCLR [52]	69.4	76.8	46.7	58.0
HiCo [57]	68.3	84.8	56.6	59.1
CMD [45]	70.6	85.4	58.3	60.9
HaLP [46]	65.8	83.6	55.8	59.0
HaLP+CMD [46]	71.0	**86.4**	59.4	61.9
DMMG [53]	**72.8**	69.9	51.5	52.3
SkeleMoCLR-	69.0	75.8	**60.6**	**63.3**

TABLE V
KNN EVALUATION RESULTS FOR DIFFERENT METHODS ON PKU-MMD.

Method	Part-I	Part-II
SkeletonCLR [5]	64.9	19.9
AimCLR [6]	73.2	19.4
SkeleMixCLR [51]	75.7	33.8
CMAL [54]	77.1	36.6
SkeleMoCLR-	**77.2**	**37.4**

Linear Evaluation. The skeletal encoder is frozen, and a fully connected layer with a Softmax activation function is added as a linear classifier for training. Compared to other methods in Tables 6 and Table 7, 2s-SkeleMoCLR- demonstrates superiority on the three datasets. On NTU-60, 2s-SkeleMoCLR- outperforms 3s-SkeAttnCLR by 1.0% and 2.0% on CS and CV protocols, respectively. On NTU-120, 2s-SkeleMoCLR- achieves accuracy levels of 77.6%/78.1% on CS and CE protocols, reaching the state-of-the-art level. On PKU-MMD, 2s-SkeleMoCLR- surpasses 3s-AimCLR by 1.3%

and 16.3% in Part-I and Part-II, respectively, with accuracies of 89.1%/54.8%.

TABLE VI
LINEAR EVALUATION RESULTS FOR DIFFERENT METHODS ON NTU-60 AND NTU-120.

Method	NTU-60		NTU-120	
	CS	CV	CS	CE
LongT GAN [25]	39.1	48.1	35.6	39.7
P&C [26]	50.7	76.3	42.7	41.7
ISC [28]	76.3	85.2	67.1	67.9
SkeletonCLR [5]	68.3	76.4	56.8	55.9
CrosSCLR-B [5]	77.3	85.1	67.1	68.6
AimCLR [6]	74.3	79.7	63.4	63.4
SkeleMixCLR [51]	79.6	84.4	67.4	69.6
SkeAttnCLR [52]	80.3	86.1	66.3	74.5
PSTL [55]	77.3	81.8	66.2	67.7
HYSP [58]	78.2	82.6	61.8	64.6
ACL [54]	78.6	84.5	68.5	71.1
CMD [45]	79.8	86.9	70.3	71.5
HiCo [57]	81.1	88.6	72.8	74.1
HaLP [46]	79.7	86.8	71.1	72.2
HaLP+CMD [46]	82.1	88.6	72.6	73.1
DMMG [53]	82.1	87.1	69.6	70.1
2s-DMMG [53]	**84.2**	89.3	72.7	72.4
3s-Colorization [50]	75.2	83.1	-	-
3s-SkeletonCLR [5]	77.8	83.4	67.9	66.7
3s-CrosSCLR-B [5]	82.1	89.2	71.6	73.4
3s-AimCLR [6]	78.9	83.8	68.2	68.8
3s-SkeleMixCLR [51]	81.0	85.6	69.1	69.9
3s-SkeAttnCLR [52]	82.0	86.5	77.1	**80.0**
3s-PSTL [55]	79.1	83.8	69.2	70.3
3s-HYSP [58]	79.1	85.2	64.5	67.3
3s-HiCLR [48]	80.4	85.5	70.0	70.4
3s-CMD [45]	84.1	**90.9**	74.7	76.1
SkeleMoCLR-	81.7	86.8	74.8	75.7
2s-SkeleMoCLR-	83.0	88.5	**77.6**	78.1

TABLE VII
LINEAR EVALUATION RESULTS FOR DIFFERENT METHODS ON PKU-MMD.

Method	Part-I	Part-II
MS^{2}L [27]	64.9	27.6
LongT GAN [25]	67.7	26.0
ISC [28]	80.9	36.0
SkeletonCLR [5]	80.9	35.2
AimCLR [6]	83.4	36.8
SkeleMixCLR [51]	89.2	51.6
ACL [54]	88.1	53.4
PSTL [55]	88.4	49.3
HiCo [57]	89.3	49.4
3s-CrosSCLR [5]	84.9	21.2
3s-AimCLR [6]	87.8	38.5
3s-SkeleMixCLR [51]	**90.6**	52.9
3s-PSTL [55]	89.2	52.3
SkeleMoCLR-	88.0	52.1
2s-SkeleMoCLR-	89.1	**54.8**

Fine-Tuning Evaluation. The skeletal encoder and linear classifier are fine-tuned using the entire dataset. The performance of 2s-SkeleMoCLR- compared to other methods on NTU-60 and NTU-120 is shown in Table 8. Experimental results demonstrate that 2s-SkeleMoCLR- is competitive with the current best methods on NTU-60. Under CS protocol, 2s-SkeleMoCLR- outperforms the state-of-the-art method (3s-Hi-TRS) by 0.6%. On NTU-120, 2s-SkeleMoCLR- achieves performance gains of 1.8% and 2.8% on CS and CE protocols, respectively, compared to 2s-DMMG. 2s-SkeleMoCLR- outperforms SkeleMoCLR-, indicating that dual-stream learning captures more action semantic information than single-stream learning.

TABLE VIII
FINE-TUNING EVALUATION RESULTS FOR DIFFERENT METHODS ON NTU-60 AND NTU-120.

Method	NTU-60		NTU-120	
	CS	CV	CS	CE
CrosSCLR [5]	86.2	92.5	80.5	80.4
AimCLR [6]	83.3	89.2	77.2	76.0
SkeleMixCLR [51]	84.5	91.1	75.1	76.0
SkeAttnCLR [52]	87.3	92.8	77.3	87.8
PSTL [55]	84.5	92.0	78.6	78.9
HYSP [58]	86.5	93.5	81.4	82.0
ACL [54]	86.9	92.8	81.7	82.7
3s-CrosSCLR [5]	86.2	92.5	80.5	80.4
3s-AimCLR [6]	86.9	92.8	80.1	80.9
3s-SkeleMixCLR [51]	87.8	93.9	81.6	81.2
3s-Colorization [50]	88.0	94.9	-	-
3s-SkeAttnCLR [52]	89.4	94.5	83.4	**92.7**
3s-PSTL [55]	87.1	93.8	81.3	82.6
2s-DMMG [53]	87.9	94.2	82.4	83.0
3s-HiCLR [48]	88.3	93.2	82.1	83.7
3s-HYSP [58]	89.1	95.2	84.5	86.3
3s-Hi-TRS [49]	90.0	**95.7**	**85.3**	87.4
SkeleMoCLR-	88.8	92.7	81.8	83.0
2s-SkeleMoCLR-	**90.6**	94.9	84.2	85.8

Transfer Learning. The transfer learning performance of the method is evaluated by training the skeletal encoder on NTU-60, NTU-120, and PKU-MMD Part-I, followed by fine-tuning with a linear layer on the PKU-MMD Part-II dataset. As shown in Table 9, our method enhances the encoder's generalization, achieving significant performance improvements. Compared to ACL, SkeleMoCLR- demonstrates accuracy improvements of 0.6%, 1.2%, and 0.7% on the three datasets, respectively. Pretraining on the NTU-120 dataset leads to substantial accuracy improvements on PKU-MMD Part-II, highlighting the benefits of the transferability of the learned representations.

TABLE IX
FINE-TUNING EVALUATION RESULTS FOR DIFFERENT METHODS ON NTU-60, NTU-120 AND PKU-MMD PART-I.

Method	NTU-60	NTU-120	PKU-MMD Part-I
LongT GAN [25]	44.8	-	43.6
MS^{2}L [27]	45.8	-	44.1
ISC [28]	45.9	52.3	45.1
MCC [59]	52.7	54.5	49.6
CrosSCLR-B [5]	54.0	52.8	-
CMD [45]	56.0	57.0	-
HiCo [57]	56.3	-	53.4
HaLP [46]	54.8	55.4	-
HaLP+CMD [46]	56.6	57.3	-
ACL [54]	61.2	62.4	58.5
SkeleMoCLR-	**61.8**	**63.6**	**59.2**

E. Qualitative Analysis

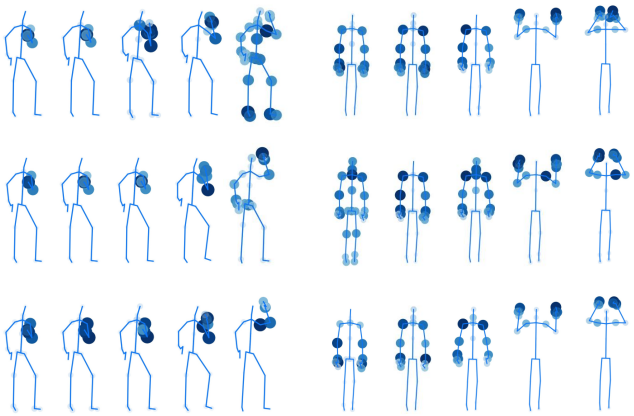
As shown in Figure 4, we employ t-SNE [60] to visualize the feature embeddings learned by SkeleMoCLR-, CrosSCLR, and AimCLR on NTU-120. To ensure a fair comparison, we randomly select 20 skeletal action categories. In comparison to competing contrastive learning methods, SkeleMoCLR- leads

to more compact feature representations for the same category and more distinguishable feature representations for different categories. This further underscores the enhanced capability of SkeleMoCLR- in learning skeletal action representations.

(a) CrosSCLR (b) AimCLR (c) SkeleMoCLR-

Fig. 4. T-SNE visualization of action features learned by (a) CrosSCLR, (b) AimCLR, and (c) SkeleMoCLR-. (Best viewed in color).

As shown in Figure 5, we visualize the output feature maps of the final layer in the 2s-SkeleMoCLR backbone on NTU-120, where circles centered around the joints represent the magnitude of the feature responses for those joints. With 5% labeled data, we compare the skeletal responses of the classic AimCLR (first row), 2s-SkeleMoCLR w/o PT pre-training (second row), and 2s-SkeleMoCLR w/ PT (third row). Compared to other rows, 2s-SkeleMoCLR w/PT can focus on the most relevant parts of the actions. For the "drinking water" action, the arm region is deeply involved, and the circles with larger radii and darker colors indicate that these joints are crucial for the skeletal action. In the "take off headphone" action, there is a more significant distribution of responses in the limb joints, with almost no response in other parts. It is evident that contrastive pre-training can enhance the model's robustness and reduce interference from noisy joints.



(a) Drink water (b) Take off headphone

Fig. 5. Feature responses of all joints in the final layer of the 2s-SkeleMoCLR backbone. The larger the response, the larger the radius of the circle centered around the joint, and the darker the color.

F. Ablation Study

With 10% labeled data, we conduct semi-supervised ablation studies using the linear evaluation protocol on NTU-60 to validate the effectiveness of different components.

Different Vision-Text models. In addition to ActionCLIP, we conduct experiments with various vision-text models during contrastive pre-training to explore their effects. X-CLIP [61] directly adapts the pre-trained image-language model for video recognition. Florence [62] further extends the CLIP approach by utilizing a unified contrastive objective. As shown in Table 10, while using more powerful vision-text models could improve our method's performance, the enhancement may not be substantial.

TABLE X
COMPARISON OF SEMI-SUPERVISED RESULTS WITH DIFFERENT VISION-TEXT MODELS WITH 10% LABELED DATA ON NTU-60 AND PKU-MMD DATASETS.

Model	NTU-60		PKU-MMD	
	CS	CV	Part-I	Part-II
X-CLIP [61]	78.1	81.6	88.1	42.1
ActionCLIP	78.4	81.8	88.3	42.1
X-Florence [61]	78.8	89.3	88.5	42.2

Different Hyperparameters. We follow the semi-supervised setting in [35] and set λ_R to 50. We further investigate the values of other hyperparameters in the loss function through ablation experiments. As shown in Table 11, the optimal values are λ_{CE} = 1 and λ_{CL} = 1, which are consistent with the default reference values for loss weights in existing semi-supervised works.

TABLE XI
COMPARISON OF SEMI-SUPERVISED RESULTS UNDER DIFFERENT HYPERPARAMETER SETTINGS WITH 10% LABELED DATA ON NTU-60 AND PKU-MMD DATASETS.

Methods	λ_{CE}	λ_{CL}	NTU-60 cs	PKU-MMD Part-I
Mean Teacher	-	-	69.1	77.6
2s-SkeleMoCLR	1	0	76.5	85.6
	1	0.1	77.0	86.3
	1	1	**78.4**	**88.3**
	1	5	77.5	87.2
	1	10	76.8	86.3
	5	1	77.2	86.7
	10	1	76.6	85.5

Different Semi-supervised Frameworks. As shown in Figure 7, we employ classical semi-supervised frameworks such as FixMatch [1], Mean Teacher [9], Noisy Student [63], and DST [64] to conduct semi-supervised experiments. As demonstrated in Table 10, contrastive pretraining result in significant improvements in various semi-supervised frameworks compared to random initialization, confirming its strong representation capabilities. Additionally, we discover that existing semi-supervised methods in the image domain are not directly suitable for the task of skeleton action recognition. Our semi-supervised framework exhibits a clear advantage in this context.

Different Components of 2s-SkeleMoCLR. In Table 11, "PT" represents contrastive pre-training, "CI" signifies category information in the queues, "RR" denotes representation regularization, and "EA" refers to extreme augmentation. The results in Table 11 indicate that, on CS and CV protocols, contrastive pre-training improves accuracy by 5.8% and 5.7%, representation regularization enhances performance by 1.2% and 0.8%. The introduction of category information in the queue increases accuracy by 3.3% and 3.8% on CS and CV

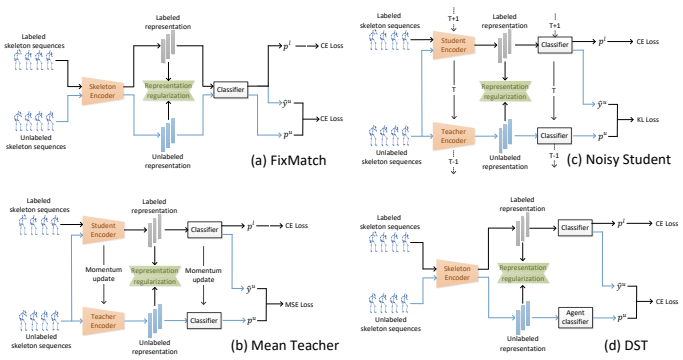


Fig. 6. Illustration of model structure based on (a) FixMatch, (b) Mean Teacher, (c) Noisy Student and (d) DST framework.

TABLE XII
COMPARISON OF SEMI-SUPERVISED RESULTS WITH DIFFERENT SEMI-SUPERVISED FRAMEWORKS ON THE NTU-60 CS PROTOCOL WITH 10% LABELED DATA.

Framework	w/o PT	w/ PT
Noisy Student	59.6	66.0
FixMatch	60.9	66.8
DST	61.4	67.7
Mean Teacher	63.0	69.1
Ours	**72.6**	**78.4**

protocols, and with the help of extreme augmentation, top-1 accuracy is improved by 2.4% and 2.7%, respectively. These results indicate that the proposed design enables the skeleton encoder to learn stronger and more robust features, making them better suited for semi-supervised tasks.

TABLE XIII
SEMI-SUPERVISED ABLATION EXPERIMENTS OF DIFFERENT COMPONENTS OF 2S-SKELEMOCLR ON NTU-60 WITH 10% LABELED DATA.

w/ PT	w/ CI	w/ RR	w/ EA	CS	CV
	✓	✓	✓	72.6	76.1
✓	✓	✓	✓	**78.4**	**81.8**
✓		✓	✓	75.1	78.0
✓	✓		✓	77.2	81.0
✓	✓	✓		76.0	79.1

G. Complexity Analysis

To analyze the model complexity, we compared different models in terms of parameters (M) and FLOPs (G). 2s-SkeleMoCLR is cleverly designed and only require skeleton data during the inference phase, which gives it a complexity advantage over other models.

V. CONCLUSION

In this paper, we propose a novel pseudo-label-based model, SkeleMoCLR, for semi-supervised skeleton action recognition. Inspired by MoCo V2 and Mean Teacher, we incorporate self-supervised models into a semi-supervised framework to generate pseudo-labels. Furthermore, we introduce category information of pseudo-labels into the memory queue of contrastive learning, making full use of the advantages of self-supervised and semi-supervised learning. Experimental results on the NTU-60, NTU-120, PKU-MMD, and NW-UCLA datasets validate the outstanding performance of SkeleMoCLR, which provides more discriminative action representations.

TABLE XIV
COMPARISON OF THE COMPLEXITY OF DIFFERENT MODELS

Models	Params(M)	FLOPs(G)
ST-GCN [19]	3.1	16.7
MS-G3D [20]	6.4	48.8
AS-GCN [21]	7.2	35.5
2s-AAGCN [65]	7.6	39.1
2s-AGCN [22]	6.9	37.3
DGNN [66]	8.1	71.1
DSTA [67]	4.1	64.7
PoseConv3D [13]	2.0	20.6
2s-MAC-Learning [33]	8.1	40.7
2s-SkeleMoCL	7.4	37.8

REFERENCES

[1] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020.
[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, pp. 1597–1607, 2020.
[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, pp. 9729–9738, 2020.
[4] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
[5] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *CVPR*, pp. 4741–4750, 2021.
[6] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *AAAI*, pp. 762–770, 2022.
[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, pp. 8748–8763, 2021.
[8] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.
[9] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017.
[10] Z. X. Bruce X B, Liu Y, "Mmnet: A model-based multimodal network for human action recognition in rgb-d videos," *TPAMI*, vol. 45, p. 3522–3538, 2023.
[11] Z. Y. Xiang W, Li C, "Generative action description prompts for skeleton-based action recognition," in *ICCV*, p. 10276–10285, 2023.
[12] L. Y. Wu H, Ma X, "Spatiotemporal multimodal learning with 3d cnns for video action recognition," *TCSVT*, vol. 32, p. 1250–1261, 2022.
[13] C. K. Duan H, Zhao Y, "Revisiting skeleton-based action recognition," in *CVPR*, p. 2969–2978, 2022.
[14] D. R. Das S, Sharma S, "Vpn: Learning video-pose embedding for activities of daily living," in *ECCV*, pp. 72–90, 2020.
[15] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *TIP*, vol. 27, no. 9, pp. 4382–4394, 2018.
[16] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *TPAMI*, vol. 41, no. 8, pp. 1963–1978, 2019.
[17] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *CVPR*, pp. 3288–3297, 2017.
[18] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3d bio-constrained skeleton model," *TIP*, vol. 28, no. 8, pp. 3959–3972, 2019.
[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.
[20] C. Z. Liu Z, Zhang H, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *CVPR*, p. 143–152, 2020.

[21] C. X. Li M, Chen S, "Actionalstructural graph convolutional networks for skeleton-based action recognition," in *CVPR*, p. 3595–3603, 2019.
[22] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, pp. 12026–12035, 2019.
[23] K. Cheng, Y. Zhang, X. He, J. Cheng, and H. Lu, "Extremely lightweight skeleton-based action recognition with shiftgcn++," *TIP*, vol. 30, pp. 7333–7348, 2021.
[24] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank, "Feedback graph convolutional network for skeleton-based action recognition," *TIP*, vol. 31, pp. 164–175, 2021.
[25] 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 *AAAI*, vol. 32, 2018.
[26] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *CVPR*, pp. 9631–9640, 2020.
[27] L. Lin, S. Song, W. Yang, and J. Liu, "Ms2l: Multi-task self-supervised learning for skeleton based action recognition," in *ACM MM*, pp. 2490–2498, 2020.
[28] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3d action representation learning," in *ACM MM*, pp. 1655–1663, 2021.
[29] P. Wang, J. Wen, C. Si, Y. Qian, and L. Wang, "Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition," *TIP*, vol. 31, pp. 6224–6238, 2022.
[30] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3d action recognition," in *ECCV*, pp. 35–51, 2020.
[31] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *TMM*, 2022.
[32] S. X. Xu B, "Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition," *arXiv preprint arXiv:2302.02327*, 2023.
[33] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *TPAMI*, 2022.
[34] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *TIP*, vol. 31, pp. 3852–3867, 2022.
[35] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou, "Adaptive consistency regularization for semi-supervised transfer learning," in *CVPR*, pp. 6923–6932, 2021.
[36] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, pp. 1010–1019, 2016.
[37] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, vol. 42, no. 10, pp. 2684–2701, 2019.
[38] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *TOMM*, vol. 16, no. 2, pp. 1–24, 2020.
[39] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, pp. 2649–2656, 2014.
[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NeurIPS Workshops*, 2017.
[41] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *ICCV*, pp. 1476–1485, 2019.
[42] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICMLW*, vol. 3, p. 896, 2013.
[43] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *TPAMI*, vol. 41, no. 8, pp. 1979–1993, 2018.
[44] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *NeurIPS*, vol. 17, 2005.
[45] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation," in *ECCV*, pp. 734–752, 2022.
[46] 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 *CVPR*, pp. 18846–18856, 2023.
[47] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *TNNLS*, 2023.
[48] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *AAAI*, vol. 37, pp. 3427–3435, 2023.
[49] Y. Chen, L. Zhao, J. Yuan, Y. Tian, Z. Xia, S. Geng, L. Han, and D. N. Metaxas, "Hierarchically self-supervised transformer for human skeleton representation learning," in *ECCV*, pp. 185–202, 2022.
[50] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3d action representation learning," in *ICCV*, pp. 13423–13433, 2021.
[51] Z. Chen, H. Liu, T. Guo, Z. Chen, P. Song, and H. Tang, "Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition," *arXiv preprint arXiv:2207.03065*, 2022.
[52] 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 *IJCAI*, 2023.
[53] S. Guan, X. Yu, W. Huang, G. Fang, and H. Lu, "Dmmg: Dual min-max games for self-supervised skeleton-based action recognition," *arXiv preprint arXiv:2302.12007*, 2023.
[54] T. Guo, M. Liu, H. Liu, W. Li, J. Guo, T. Wang, and Y. Li, "Joint adversarial and collaborative learning for self-supervised action recognition," *arXiv preprint arXiv:2307.07791*, 2023.
[55] Y. Zhou, H. Duan, A. Rao, B. Su, and J. Wang, "Self-supervised action representation learning from partial spatio-temporal skeleton sequences," in *AAAI*, 2023.
[56] U. P. K. Kundu J N, Gor M, "Unsupervised feature learning of human actions as trajectories in pose embedding manifold," in *WACV*, p. 1459–1467, 2019.
[57] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang, "Hierarchical contrast for unsupervised skeleton-based action representation learning," in *AAAI*, vol. 37, pp. 525–533, 2023.
[58] L. Franco, P. Mandica, B. Munjal, and F. Galasso, "Hyperbolic self-paced learning for self-supervised skeleton-based action representations," in *ICLR*, 2023.
[59] Y. Su, G. Lin, and Q. Wu, "Self-supervised 3d skeleton action representation learning with motion consistency and continuity," in *ICCV*, pp. 13328–13338, 2021.
[60] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[61] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *ECCV*, pp. 1–18, 2022.
[62] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.
[63] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, pp. 10687–10698, 2020.
[64] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," in *NeurIPS*, 2022.
[65] C. J. Shi L, Zhang Y, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *TIP*, vol. 29, p. 9532–9545, 2020.
[66] C. J. Shi L, Zhang Y, "Skeleton-based action recognition with directed graph neural networks," in *CVPR*, p. 7912–7921, 2019.
[67] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *ACCV*, 2020.