Cross-Modal Contrastive Pre-training for Few-Shot Skeleton Action Recognition

Mingqi Lu, Siyuan Yang, Xiaobo Lu and Jun Liu

***Abstract*—This paper proposes a novel approach for few-shot skeleton action recognition that comprises of two stages: cross-modal pre-training of a skeleton encoder, followed by fine-tuning of a cosine classifier on the support set. The pre-training and fine-tuning approach has been demonstrated to be more effective for handling few-shot tasks compared to utilizing more intricate meta-learning methods. However, its success relies on the availability of a large-scale training dataset, which yet is difficult to obtain. To address this challenge, we introduce a cross-modal pre-training framework based on Bootstrap Your Own Latent (BYOL), which considers skeleton sequences and their corresponding videos as augmented views of the same action in different modalities. By utilizing a simple regression loss, the framework is able to transfer robust and high-quality vision-language representations to the skeleton encoder. This allows the skeleton encoder to gain a comprehensive understanding of action sequences and benefit from the prior knowledge obtained from a vision-language pre-trained model. The representation transfer enhances the feature extraction capability of the skeleton encoder, compensating for the lack of large-scale skeleton datasets. Extensive experiments on the NTU RGB+D, NTU RGB+D 120, PKU-MMD, NW-UCLA, and MSR Action Pairs datasets demonstrate that our proposed approach achieves state-of-the-art performances for few-shot skeleton action recognition.**

***Index Terms*—Few-shot skeleton action recognition, contrastive learning, knowledge distillation.**

I. Introduction

Skeleton-based action recognition has gained widespread popularity due to its advantages in terms of computational efficiency, robustness, and privacy protection. Current skeleton-based action recognition algorithms [1]–[5] mainly focus on the many-shot classification problem, where multiple labeled training samples are available for each category. However, the acquisition of skeleton sequences is not as convenient as that of images, and requires the utilization of depth cameras or pose estimation algorithms. In low-data scenarios, the few-shot approaches show great potential for both development and practical applications. While the field of few-shot skeleton action recognition is still in its early stages, previous research has primarily adopted classic few-shot algorithms, such as ProtoNet [6] for images and DTW [7] for videos. Skeleton-DML [8] transforms skeleton data into images, while the DTW-based JEANIE [9] addresses view alignment challenges. The recent DASTM [10] method employs Soft-DTW [11] for temporal alignment and still relies on ProtoNet to assess the similarity between query samples and prototypes. Our approach departs from prior methods by attempting to tackle the challenge of few-shot skeleton action recognition in a straightforward manner.

In this paper, we introduce a novel "Pre-training and Fine-tuning" approach for few-shot skeleton action recognition. Similar to the ImageNet [12] pre-trained model in the image domain and BERT [13] in the natural language processing domain, pre-trained encoders have a vital function in feature extraction in this approach. However, there is currently a lack of substantial pre-trained models in the skeleton-based action recognition field. Considering the success of vision-language pre-training models [14], we explore the feasibility of utilizing the vision-language contrastive learning training scheme to skeleton encoders. Nevertheless, this is not a direct option due to the scarcity of a large-scale dataset containing skeleton-text pairs and the high cost of training a skeleton-language model (which requires significant GPU resources). Moreover, directly transferring representational knowledge from images to skeletons is challenging due to the sparse and non-uniform distribution of skeleton data. To address these challenges, we propose a novel cross-modal contrastive pre-training framework that exploits a vision-language pre-trained model to transfer high-quality representations to skeleton encoders. Skeleton data, whether acquired through a depth camera or a pose estimation algorithm, is usually accompanied by RGB video that depicts the same action and is closely related. Inspired by image-based self-supervised learning, we view the video as an augmentation of the skeleton over different modalities.

Self-supervised learning is a technique of representation learning that leverages unlabeled data. The state-of-the-art contrastive approaches [15]–[17] seek to minimize the dissimilarity between representations of the same image under different augmentations and maximize the difference between representations of different images.

Mainstream methods in self-supervised learning aim to

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).

Siyuan Yang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore (email: siyuan005@e.ntu.edu.sg).

Jun Liu is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore, and also with Lancaster University, UK (email: jun_liu@sutd.edu.sg).

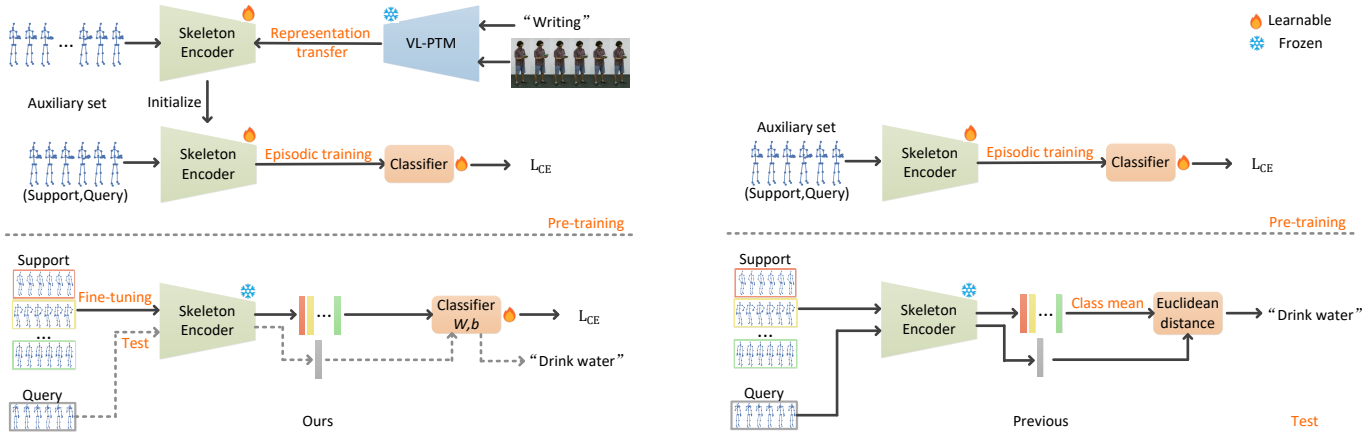


Fig. 1. The overall architecture of our model. Given the scarcity of large-scale skeleton pre-training datasets, we adopt a two-stage approach to pre-train the skeleton encoders using a vision-language pre-trained model (VL-PTM). The two stages involved are representation transfer and episode training. After pre-training, the skeleton encoder remains fixed, and the weights (W) and biases (b) of classifier are fine-tuned using the support samples as training data. Compare the right figure to highlight the innovations in our approach.

optimize the transfer of instructor expertise to students by enforcing alignment between the final embeddings of the student and teacher models. Some self-supervised techniques for representation learning use knowledge distillation to boost the efficacy of student models. SimCLR-V2 [18] employs logits during fine-tuning to convey task-specific knowledge. SEED [19] mimics the distribution of similarity scores between teacher and student models using a dynamically maintained queue. However, SEED heavily depends on the MoCo [16] framework, suggesting the continuous need to maintain the memory bank during distillation. In contrast, the Bootstrap Your Own Latent (BYOL) approach [20] closely adheres to knowledge distillation, reducing the gap between teacher and student representations. BYOL trains an online network to predict the representations of the same image under different augmentations, without the requirement of explicit negative samples. Inspired by this, we present a cross-modal representation transfer framework, which utilizes pairs of skeleton sequences and RGB videos as two augmented views of the same action. The skeleton encoder learns to identify distinctive features by predicting the embeddings generated by a pre-trained vision-language model. Our framework is motivated by the need for few-shot learning, which attempts to learn instead of merely recognizing samples in the training set. The vision-language pre-trained model provides a comprehensive understanding of skeleton sequences, and we leverage this knowledge for representation transfer, enabling the skeleton encoder to extract features more effectively and distinguish between samples. BYOL is chosen as the preferred method, as it can learn modality-independent representations with a simple regression loss.

The main contributions of this paper are: (1) a novel cross-modal contrastive pre-training approach for few-shot skeleton action recognition; (2) a representation transfer framework that leverages vision-language pre-trained models for scenarios lacking large-scale pre-training datasets; (3) our proposed approach outperforms the state-of-the-art on NTU RGB+D, NTU RGB+D 120, PKU-MMD, NW-UCLA, and MSR Action Pairs datasets.

II. Related Work

A. Skeleton-based Action Recognition

Inspired by the observation that the human skeleton is naturally a topological graph, GCNs have attracted increasing attention in skeleton-based action recognition. ST-GCN [4] leveraged spatiotemporal GCN to capture human joint relationships in both spatial and temporal dimensions. Building upon ST-GCN, Shi *et al*. proposed 2s-AGCN [5], which is capable of dynamically learning graph topology in an end-to-end manner. CTR-GCN [3] employed refined spatial attention in the channel dimension to learn the dynamic features of different channels. Shao et al. [21] utilized a multi-stream neural network for cross-view action recognition from skeleton data. MCMT-Net [22] captured the relationships within skeleton sequences through an efficient decomposition of spatio-temporal graphical models. Recently, self-supervised skeleton-based action recognition has emerged as a promising direction. Yang *et al*. [23] proposed representing skeleton sequences as skeleton clouds to learn their spatial and temporal information by solving the skeleton cloud coloring problem. CrosSCLR [24] learned skeleton sequence representations using a momentum contrast framework, while AimCLR [25] built upon CrosSCLR and incorporates an energy-based attention-guided dropout module and nearest neighbor mining. Moliner *et al*. [26] applied BYOL to skeleton-based action recognition using two very different pipelines of conservative and aggressive enhancements. In contrast to existing approaches, our approach utilizes the representation transfer framework to extract discriminative features from vision-language pre-trained models, thereby facilitating skeleton representation learning. This unique approach sets our approach apart from others in the field.

B. Few-Shot Learning

Few-shot learning is a technique that uses a limited amount of labeled data to classify query samples. In the field of few-

shot learning, meta-learning has been the dominant approach, with a majority of studies focusing on image classification [27]–[30]. ProtoNet [6] computed the distance between samples and class prototypes, while FEAT [31] defined a set-to-set transformation for learning task-specific feature embeddings. Simon *et al*. [32] used a subspace approach as the central block of a dynamic classifier. MSML [33] extracted multi-scale features and learns multi-scale relationships between samples, and Zhang *et al*. [34] optimized the model by formulating it as a variational inference problem Recently, researchers have shown that embedding models pre-trained on a large pre-training dataset can achieve comparable results to many state-of-the-art meta-learning algorithms [35]–[38]. Few-shot action recognition has also become a topic of interest due to the need to identify emerging new actions [39]–[43]. Several works [44]–[47] have adopted the ProtoNet scheme to compute the similarity between video samples. However, these methods cannot be directly applied to few-shot skeleton action recognition tasks since they rely on RGB images or videos with richer data meaning than skeleton sequences. Therefore, our work takes a different approach, starting with the simple method of pre-training and fine-tuning, and attempting to solve the few-shot skeleton action recognition problem through a few-shot learning baseline. However, the lack of a large-scale skeleton dataset presents a significant challenge for pre-training. To address this issue, we propose a representation transfer framework that leverages the power of large-scale vision-language pretrained models.

C. Few-Shot Skeleton Action Recognition

Leveraging the NTU RGB+D 120 dataset, Liu *et al.* [48] first presented an Action-Part Semantic-Relevance aware (APSR) approach for one-shot skeleton action recognition. Sabater *et al*. [49] employed a Temporal Convolutional Network (TCN) to extract skeleton features and calculates the cosine similarity between query and support features. Memmesheimer *et al*. [8], [50] transformed the skeleton sequence into images and performs classification based on metric learning. JEANIE [9] proposed the temporal and viewpoint alignment of support and query samples. SMAM [51] proposed an adaptive matching module for similarity measure. The most recent work DASTM [10] adopted the Prototypical Networks architecture and performs spatial alignment through rank-maximization and temporal alignment based on DTW. According to DASTM, compared to ProtoNet-based methods, meta-learning methods such as MAML [28] required large memory overhead to memorize multiple gradient steps and cannot effectively incorporate larger encoders like ST-GCN.

III. Problem Definition

This paper addresses the challenge of few-shot skeleton-based action recognition. The skeleton data set $\mathrm{D} = \{(x_i, y_i \mid y_i \in C_D)\}$ contains N_D categories, which are divided by category into a base set $\mathrm{D_B} = \{(x_i, y_i \mid y_i \in C_B)\}$ and a novel set $\mathrm{D_N} = \{(x_i, y_i \mid y_i \in C_N)\}$, where $\mathrm{C_B} \cap \mathrm{C_N} = \varnothing$ and $\mathrm{C_B} \cup \mathrm{C_N} = \mathrm{C_D}$. The novel set $\mathrm{D_N}$ is composed of N categories, with each category containing K labeled sample, forming the support set S, in a K-shot scenario. Our approach involves training the skeleton encoder on the base set $\mathrm{D_B}$ using a vision-language pre-trained model and fine-tuning the classifier on the support set S.

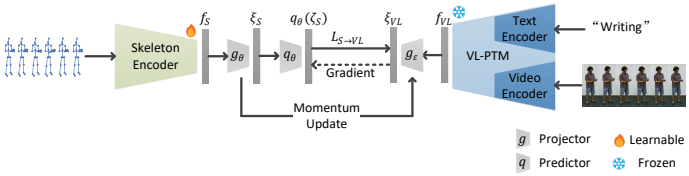


Fig. 2. Illustration of the representation transfer framework

IV. Approach

As there is a limited number of labeled novel set samples, the pre-training and fine-tuning manners in our approach should be different from that in traditional transfer learning. Due to the scarcity of large-scale skeleton-text pre-training datasets, we adopt a two-stage approach to pre-train the skeleton encoders by leveraging a pre-trained vision-language model. These two stages consist of representation transfer and episode training, as shown in Figure 1. Once the pre-training is complete, the skeleton encoder remains fixed, and the support samples are used as training data, with only the weights W and biases b in the classifier being fine-tuned. During forward propagation, the model calculates the cosine similarity between the support and query samples to perform classification.

A. Vision-Language Representation

As shown in Figure 2, we leverage the X-CLIP [52] as the video-text embedding model. X-CLIP [52] is a video-text pre-training model based on CLIP [14], comprising of a video encoder and a text encoder, which creates dense connections between text and video through inner product contrast, demonstrating that the features of both modalities are well aligned in the same feature embedding space. The embedded video features and text features are represented as f_v and f_t respectively, and the vision-language features f_{VL} are obtained as $f_{VL} = f_v + f_t$. This pre-trained vision-language knowledge is then transferred to the skeleton encoder through our proposed framework. During training, the parameters of the X-CLIP encoders are frozen while the skeleton encoder is made learnable.

B. Pre-training of Skeleton Encoder

We introduce a representation transfer framework for pre-training the skeleton encoder. Our methodology is built on BYOL [20], which employs two views of the same input to train an online network to predict the representation of a target network. However, our approach differs from BYOL in several crucial aspects. As illustrated in Figure 2, the representation transfer framework consists of two networks: the skeleton network and the vision-language network, with the latter providing the regression target for training the former. Specifically, we treat skeleton and vision-language information

as two separate views and transfer representation from a pre-trained vision-language model to a skeleton encoder. The skeleton encoder E_S extracts features $f_S = E_S(x)$ from a skeleton sequence x, while the vision-language pre-trained model E_{VL} is used to extract vision-language features f_{VL}. The parameters of the vision-language pre-trained model are frozen during representation transfer. Projectors g_θ and g_ε are defined to obtain skeleton embeddings $\zeta_S = g_\theta(f_S)$ and vision-language embeddings $\zeta_{VL} = g_\varepsilon(f_{VL})$, which are used by the predictor q_θ to make predictions $q_\theta(\zeta_S)$ and regress the vision-language embedding ζ_{VL} using the loss $L_{S\to VL}$:

$$L_{S\to VL} = \left\| \overline{q_\theta(\zeta_S)} - \overline{\zeta_{VL}} \right\|_2^2 = 2 - 2 \cdot \frac{< q_\theta(\zeta_S), \zeta_{VL} >}{\|q_\theta(\zeta_S)\|_2 \cdot \|\zeta_{VL}\|_2} \tag{1}$$

In the iterative training phase, the gradients are only back-propagated through the skeleton network, and the vision-language projector is updated using the exponential moving average of the skeleton projector weights:

$$\xi \leftarrow \lambda\xi + (1-\lambda)\theta, \lambda \in [0,1]. \tag{2}$$

Our approach differs from BYOL in that it adopts a unidirectional regression learning from skeleton to vision-language. The Multi-Layer Perceptron (MLP) architecture used in our work is identical to the one used in the original BYOL and consists of a linear layer with an output size of 4096, followed by batch normalization, a rectified linear unit, and a final linear layer that generates a 256-dimensional embedding. Our experiments show that the incorporation of predictors into the skeleton network and the utilization of moving averages for updating the vision-language projector leads to an increased encoding of vision-language information within the skeleton projections. After representation transfer, only the skeleton encoder is retained and its weights are initialized using the learned weights from the transfer process. The skeleton encoder is then further trained through episodic training on a skeleton base set, which is a conventional meta-learning training method for few-shot tasks that minimize covariate shifting.

C. Fine-tuning of Cosine Classifier

After pre-training the skeleton encoder on the base set, we fine-tune the cosine classifier using the support set to enhance performance. For each sample (x_j, y_j) in the support set, where $1 \le j \le N$, the feature vector $E_S(x_j)$ is extracted from the pre-trained skeleton encoder E_S and fed into the softmax classifier. The resulting output probability distribution is represented as $p_j = \mathrm{softmax}(WE_S(x_j) + b)$, where W and b represent the weights and biases of the softmax classifier, respectively.

Given the limited sample size in the support set, we employ a weight initialization method from Tian *et al.* [38] instead of utilizing a random initialization. The weights W are initialized as $\mathrm{M} = [E_S(x_1), \cdots, E_S(x_i), \cdots, E_S(x_N)]^T$, and the biases b are initialized as a zero vector. In this case, the elements of the classifier output vector p_j reflect the degree of similarity between the features q of the query sample and the features of each individual class in the support set. Additionally, we employ the cosine similarity $\mathrm{sim}_{\cos}(w_j, q)$ between w_j and q in the softmax classifier, which forms a cosine classifier, to minimize the intra-class variance among features:

$$\mathrm{p} = \mathrm{softmax}\left(\begin{bmatrix} \mathrm{sim}_{\cos}(w_1, q) + b_1 \\ \cdots \\ \mathrm{sim}_{\cos}(w_j, q) + b_j \\ \cdots \\ \mathrm{sim}_{\cos}(w_N, q) + b_N \end{bmatrix} \right), \tag{3}$$

where $\mathrm{sim}_{\cos}(w_j, q) = \frac{w_j^T q}{\|w_j\|_2 \|q\|_2}$.

The objective of the fine-tuning stage is to optimize the parameters W and b of the softmax classifier on the support set. For each sample (x_j, y_j) in the support set, we calculate the Cross-Entropy of the output p_j and the label y_j. The mean of the Entropy of the output p, H(p), is also calculated and added to the fine-tuning cross-entropy loss as an entropy regularization to prevent overfitting:

$$L_{\text{fine-tune}} = L_{CE} + H(p) = L_{CE} - \sum_{j=1}^{N} p_j \log p_j. \tag{4}$$

V. Experiments

A. Dataset and Evaluation Protocol

We conduct experiments on five public datasets: NTU RGB + D 60 [53], NTU RGB + D 120 [48], PKU-MMD [54], NW-UCLA [55], and MSR Action Pairs [56]. More details about the exemplars are provided in the appendix.

NTU RGB+D 60/120. NTU RGB+D 60 [53] is a skeleton-based action recognition dataset with 60 different categories, including daily actions, interactive actions, and health-related actions. The dataset comprises 56,880 video samples performed by 40 different subjects, and each body skeleton contains 3D coordinates for 25 joints. NTU-120 [48] is an expansion of NTU-60 that consists of 120 action classes (daily/health-related) and 114,480 RGB + D video samples taken with 106 unique human participants. According to its official protocol [48], NTU-120 is split into a 100-class auxiliary set and a 20-class evaluation set with non-overlapping classes. Each class in the evaluation set contains only one reference sample serving as the exemplar. We employ the 100-class auxiliary set to train the models for general performance assessment. For the auxiliary reduction experiment, aligning with the benchmarks in [48], we apply a variable control on the auxiliary class size range of 20, 40, 60, 80, and 100. Following dataset splitting in NTU-120 [48], we introduce a one-shot skeleton action recognition setting on NTU-60. The one-shot evaluation set includes 10 novel classes whereas the auxiliary set has 50 classes. A1, A7, A13, A19, A25, A31, A37, A43, A49 and A55 are chosen as novel classes.

PKU-MMD [54] is a benchmark dataset widely used for human behavior analysis, comprising two subsets. In our study, we select the first subset, PKU-MMD Part-I, which consists of 1,076 untrimmed video sequences with 51 action categories performed by 66 subjects. After removing 6 invalid samples without skeletal frames, we obtain 21,545 valid action sequences with 51 annotated action categories. For our experiments, we choose 11 action categories for testing and

40 for training. The test categories are A01, A06, A11, A16, A21, A26, A31, A36, A41, A46, and A51. To perform 1-shot and 5-shot testing, we select one or five samples, respectively, from each test category to serve as the exemplar.

NW-UCLA dataset [55] comprises 1,494 video clips covering 10 action categories. Each action is performed by 10 different subjects. We randomly select 5 action classes for training and the remaining classes for testing in the NW-UCLA dataset. Due to the dataset's limited size, we perform 20 train/test splits to minimize errors. We adopt the 1-shot and 5-shot evaluation protocols, averaging the results over all 20 splits.

MSR Action Pairs [56] is a dataset containing 360 video clips of 6 action pairs (12 action categories), including pick up/put down box, lift/place box, push/pull chair, wear/take off hat, put on/take off backpack, and stick/remove poster. Each action pair has very similar motion trajectories, and the actions are performed three times by 10 subjects, resulting in 353 activity samples. We randomly select 3 action pairs for the training set (6 classes) and the remaining 3 action pairs (6 classes) for the test set, resulting in 20 different train/test split combinations. The use of action pairs in our train/test split enables us to test whether the algorithm can classify unseen action pairs with similar motion trajectories. We evaluate our model using the 1-shot and 5-shot protocols, averaging over all 20 splits.

B. Implementation Details

The selected skeleton encoder for this study is AGCN [5], with data preprocessing techniques aligned with CTR-GCN [3]. The default vision-language pre-trained model is X-CLIP [52] (ViT-B/16 [57]).We sparsely sample 8 frames from the RGB videos corresponding to each segment in the skeleton dataset and perform spatial cropping based on human joint coordinates. In the representation transfer framework, we only train the skeleton encoder, projectors, and predictor. We use the Adam optimizer with an initial learning rate of 0.1 and a target decay rate of 0.99 for 100 epochs of training. The batch size is set to 128, with a weight decay of 0.01. Specifically, for the predictor, its learning rate is multiplied by 10. During this process, the vision-language encoders remain frozen. Episodic training is used to continue training the skeleton encoder connected to a cosine classifier on the skeleton base set, with training data composed into episodes. For the N-way K-shot setting, N classes are randomly sampled, each class with K examples, to form the support set, and each sample in the query set belongs to one of the N classes. SGD with a momentum of 0.9 and weight decay of 0.0001 is used as the optimizer. The initial learning rate is set to 0.001 and the cosine annealing strategy is employed. The cosine classifier is fine-tuned on the support set for 20 epochs, and the initial learning rate is set to 0.0005. All experiments are conducted on a GTX 3090 GPU.

C. Comparison Results

We evaluate our approach against several competing methods on five benchmark datasets, namely NTU RGB + D 120, NTU RGB + D 60, PKU-MMD, NW-UCLA, and MSR Action Pairs, as presented in Tables I, II, III, IV, V, and VI. Among these methods are SL-DML [50] and Skeleton-DML [8], which are metric learning-based approaches that transform skeleton data into images. SMAM [51] introduces an adaptive module that utilizes a metric matching mechanism. DASTM [10] uses the prototype network as the basic few-shot solution and employs Soft-DTW [11] to align the skeleton sequences. Additionally, we compare our approach with several image-based few-shot methods, including ProtoNet [6], FEAT [31], and Subspace [32]. All the skeleton encoders in the comparison default to AGCN. As shown in Tables I, II, III, IV, V and VI, our proposed approach outperforms the state-of-the-art methods by a large margin on all five datasets, indicating the effectiveness of the proposed method on few-shot skeleton action recognition. Note that, the one-shot/five-shot experimental results for [8], [50] on PKU-MMD and five-shot performances for [8], [49], [50], [58], [59] on NTU RGB+D and NTU RGB+D 120 are from SMAM [51].

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE NTU-120 DATASET

Method	1-shot	5-shot
Attention [58]	41.0	-
Fully Connected [58]	42.1	52.4
Average Pooling [59]	42.9	51.1
APSR [48]	45.3	-
TCN-Oneshot [49]	46.5	60.3
SL-DML [50]	50.9	64.0
Skeleton-DML [8]	54.2	65.5
JEANIE [9]	57.0	-
Soft-DTW [11]	56.9	67.1
ProtoNet [6]	58.8	68.6
FEAT [31]	58.1	66.9
Subspace [32]	59.7	69.8
SMAM [51]	56.4	65.9
DASTM [10]	60.1	69.3
Ours(AGCN)	70.4	79.5
Ours(CTR-GCN)	**71.0**	**79.7**

TABLE II
EVALUATION OF DIFFERENT TRAINING SET SIZES ON THE NTU-120 DATASET

Training Classes	20	40	60	80	100
APSR [48]	29.1	34.8	39.2	42.8	45.3
SL-DML [50]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [8]	28.6	37.5	48.6	48.0	54.2
JEANIE [9]	38.5	44.1	50.3	51.2	57.0
Soft-DTW [11]	34.5	42.9	48.6	51.4	56.9
ProtoNet [6]	35.8	44.5	50.7	52.1	58.8
FEAT [31]	37.1	44.1	47.2	53.7	58.1
Subspace [32]	36.4	43.9	52.9	54.1	59.7
SMAM [51]	35.8	46.2	51.7	52.2	56.4
DASTM [10]	36.5	43.1	51.0	53.9	60.1
Ours(AGCN)	**43.3**	**54.5**	**62.1**	65.7	70.4
Ours(CTR-GCN)	43.0	**54.5**	61.9	**66.0**	**71.0**

D. Model Analysis

Attribution Visualization. To validate the efficacy of our approach, we present the results of ProtoNet, DASTM, and our

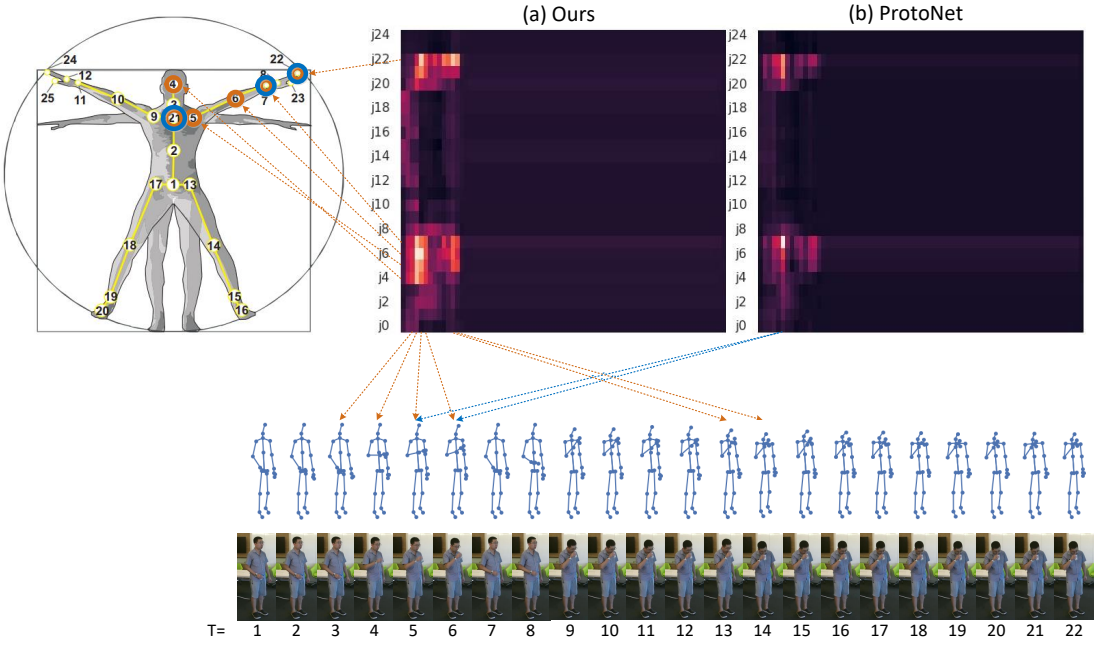


Fig. 3. Illustration of Joint-frame heatmaps for the "drink water" action. The RGB frames reveal that the most prominent features of the "drink water" action are present in frames 3, 4, 5, 6, 13, and 14. However, ProtoNet only considers the initial few frames and neglects the equally crucial frames 13 and 14. In contrast, our approach can effectively identify the two most critical moments in the action, as demonstrated by the clear hand movement captured in frames 13 and 14. In the spatial dimension, our approach precisely focuses on all the skeleton points relevant to the action of drinking, including the (4) head, (5) left shoulder, (6) left elbow, (7) left wrist, (21) spine, and (22) tip of the left hand, thereby offering a more comprehensive representation of the action (best viewed in color).

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE NTU-60 DATASET.

Method	1-shot	5-shot
Fully Connected [58]	60.9	64.2
Average Pooling [59]	59.8	61.2
TCN-Oneshot [49]	64.8	66.8
SL-DML [50]	71.4	77.0
Skeleton-DML [8]	71.8	77.6
Soft-DTW [11]	72.8	79.4
ProtoNet [6]	74.7	81.1
FEAT [31]	73.4	78.9
Subspace [32]	75.6	82.6
SMAM [51]	73.6	79.0
DASTM [10]	76.9	83.1
Ours(AGCN)	**83.9**	89.8
Ours(CTR-GCN)	83.7	**90.0**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE PKU-MMD DATASET.

Method	1-shot	5-shot
SL-DML [50]	67.0	73.0
Skeleton-DML [8]	68.6	73.7
Soft-DTW [11]	73.5	80.0
ProtoNet [6]	77.2	83.3
FEAT [31]	74.6	80.3
Subspace [32]	75.1	82.2
SMAM [51]	70.4	74.2
Ours(AGCN)	**87.3**	**92.8**
Ours(CTR-GCN)	86.7	92.3

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE NW-UCLA DATASET.

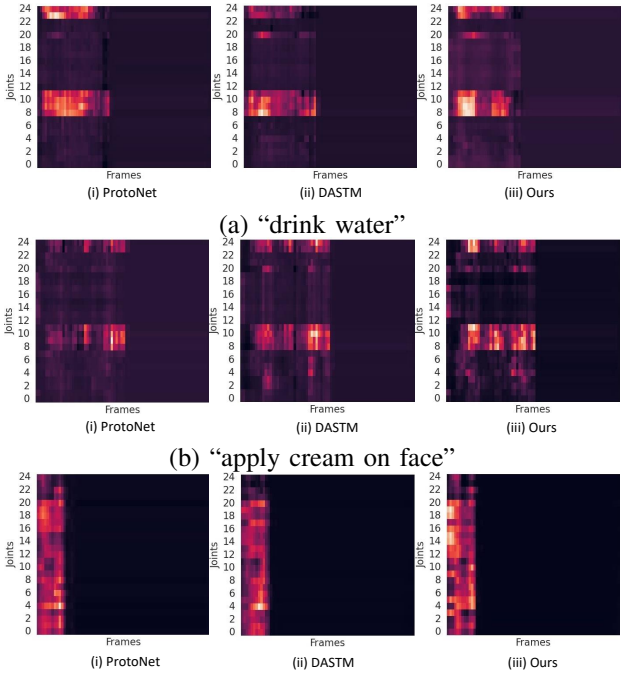
Method	1-shot	5-shot
SL-DML [50]	66.2	76.1
Skeleton-DML [8]	69.6	77.9
Soft-DTW [11]	71.1	81.0
ProtoNet [6]	73.6	81.6
FEAT [31]	71.7	80.2
Subspace [32]	72.4	81.0
Ours(AGCN)	82.1	88.9
Ours(CTR-GCN)	**82.5**	**89.5**

approach on NTU-120 using visualization techniques, which enable a more comprehensive examination of the learned representations. We employ the attribution algorithm BIG [60] to interpret the model predictions, and the resulting attribution scores are presented as joint-frame heatmaps, as depicted in Figure 3 and 4. Generally, higher scores indicate the stronger relevance of the corresponding input features to the prediction. The brightness of the grids in the heatmap corresponds to their level of importance in the prediction. Compared to prior previous methods, BIG provides a more precise interpretation while also mitigating the issue of baseline sensitivity. As depicted in Figure 4, the learned representations in our approach demonstrate a higher degree of discriminative ability compared to those generated by ProtoNet and DASTM. This observation indicates the superior capacity of our approach in the realm of few-shot skeletal action representation learning.

Failure Cases. Figure 5 illustrates instances of failure in recognizing skeleton actions using the proposed method, namely A067 "hush" and A103 "yawn". Additionally, A001 "drink water" and A085 "apply cream on face", as well as A055

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE MSR ACTION PAIRS DATASET.

Method	1-shot	5-shot
SL-DML [50]	70.1	76.2
Skeleton-DML [8]	71.7	78.7
Soft-DTW [11]	73.8	80.6
ProtoNet [6]	79.3	84.9
FEAT [31]	75.4	82.1
Subspace [32]	74.5	81.0
Ours(AGCN)	**86.2**	**91.1**
Ours(CTR-GCN)	86.0	87.7



(a) "drink water"

(b) "apply cream on face"

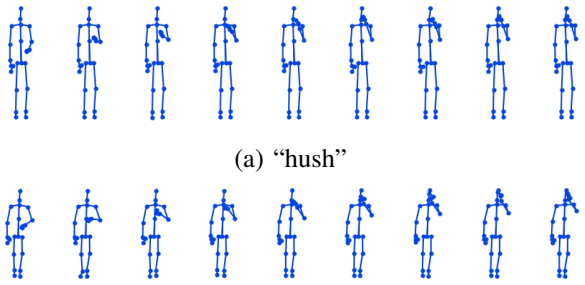
(c) "grab other person's staff" (second person).

Fig. 4. Joint-frame heatmaps for (a) "drink water", (b) "apply cream on face" and (c)"grab other person's staff"(best viewed in color).

"hugging" and A109 "grab stuff", are also prone to mutual confusion. We analyze the reasons for these prediction failures to be the lack of dense spatial cues from RGB images, as only sparse skeleton modalities are utilized, resulting in significant similarities between these actions. Taking Figure 5 as an example, the nearly identical trajectories of the left and right hands make it challenging to distinguish between "hush" and "yawn".

Different Pre-training Strategies. The pre-training phase plays a crucial role in extracting discriminative features for few-shot skeleton action recognition from sparse skeleton data. In this section, we delve into the pre-training strategy and investigate five additional methods in addition to the proposed representation transfer framework, as follows:

"Ours (traditional training)" only performs batch training on the base class skeleton data; "Ours (episodic training)" performs episodic pre-training on the base class skeleton data; "Skeleton_BYOL" pre-trains on unlabeled base class skeleton data using Moliner *et al*. [26] method without a meta-learning stage; "Skeleton_BYOL-Meta" and "AimCLR-Meta",



(a) "hush"

(b) "yawn"

Fig. 5. Failure cases for few-shot skeleton action recognition (a) "hush" and (b)"yawn"(best viewed in color).

TABLE VII
COMPARISON OF DIFFERENT PRE-TRAINING STRATEGIES USING ONE-SHOT EVALUATION PROTOCOL.

Training Strategy	NTU-120	NTU-60	PKU-MMD
w/o cross-modal pre-training			
Soft-DTW [11]	56.9	72.8	73.5
ProtoNet [6]	58.8	74.7	77.2
FEAT [31]	58.1	73.4	74.6
Subspace [32]	59.7	75.6	75.1
Skeleton_BYOL [26]	50.6	67.8	70.0
Skeleton_BYOL-Meta	62.1	74.7	78.8
AimCLR-Meta [25]	64.3	76.4	80.4
Ours(traditional training)	53.4	70.2	72.1
Ours(episodic training)	57.0	73.1	74.9
w/ cross-modal pre-training			
Soft-DTW [11]	64.0	78.7	80.3
ProtoNet [6]	66.5	80.2	84.3
FEAT [31]	66.8	79.7	82.7
Subspace [32]	67.8	82.4	83.0
Ours	**70.4**	**83.9**	**87.3**

respectively, represent using Moliner *et al*. [26] and AimCLR [25] to conduct self-supervised learning on unlabeled base class skeleton data before episodic training.

We also compare the performance of classic algorithms, including Soft-DTW, ProtoNet, FEAT, and Subspace, on three datasets under two conditions: with and without cross-modal pre-training. The term "w/ cross-modal pre-training" denotes the usage of the cross-modal pre-trained skeleton encoder as the initialization for the encoders of each algorithm. It is evident that our proposed pre-training framework significantly enhances the classification performance for different algorithms, showcasing its universality.

As shown in Table VII, our proposed representation transfer framework for pre-training outperforms these pre-training strategies significantly. Although self-supervised learning is only performed on unlabeled skeleton data, it brings considerable gains to skeleton representation learning. Results in Table 7 demonstrate the effectiveness of our proposed representation transfer framework.

Different Representations. We evaluate the impact of three different representations (f_t: text-only, f_v : video-only, and $f_t + f_v$: video-text) from the vision-language pre-trained model, and the comparison results are shown in Table VIII. It is worth noting that we only utilize the vision-language pre-trained model during the pre-training phase. The experimental results demonstrate that the skeleton encoder can benefit from

TABLE VIII
COMPARISON OF DIFFERENT REPRESENTATIONS IN TRANSFER USING ONE-SHOT EVALUATION PROTOCOL.

Representation	NTU-120	NTU-60	PKU-MMD
None	58.0	74.1	75.9
f_t	62.3	77.4	79.3
f_v	67.1	80.8	84.9
$f_t + f_v$	**70.4**	**83.9**	**87.3**

TABLE IX
COMPARISON OF DIFFERENT INITIALIZATIONS USING ONE-SHOT EVALUATION PROTOCOL.

Initialization	NTU-120	NTU-60	PKU-MMD
Random	41.6	60.1	58.8
ImageNet pre-trained	54.2	71.8	68.6
Representation Transfer	**61.3**	**75.0**	**71.2**

both visual and linguistic representation sources. Unlike labels, videos have a one-to-one correspondence with skeleton sequences, indicating the dominant role of visual representations in the transfer learning framework. Moreover, the textual features generated by the vision-language pre-trained model complement and enhance the overall representations.

Different Initialization. We conduct ablation studies on different Initialization for ResNet-18 [61] based Skeleton-DML. Skeleton-DML [8] transforms skeleton sequences into images. By default, ResNet-18 is initialized with the ImageNet pre-trained model. As shown in Table IX, the final accuracy of the randomly initialized ResNet-18 is only 41.6%, significantly lower than the results obtained from both the ImageNet pre-trained model initialization and the representation transfer initialization. These differences in performance emphasize the importance of pre-training and demonstrate that our representation transfer framework is particularly valuable in the absence of large skeleton datasets. While representation transfer significantly improves the performance of the ResNet-based skeleton encoder, the transformation of 3D skeleton sequences into 2D image results will lose the important spatial depth information, leading to a decline in performance when compared to GCN-based models.

Framework Design. We delve deeper into the structural design of the representation transfer framework. Specifically, we examine the impact of removing the moving average target, which sets the target decay rate to 0. Our findings, as presented in Table X, demonstrate that the target weight and online weight instant update result in a slight degradation in performance compared to the default setting. This suggests that the representation transfer framework still relies on the guided behavior of BYOL to learn skeleton representations. Furthermore, we investigate the impact of eliminating the predictor from the framework, which is incorporated in BYOL to prevent network collapse. The results in Table X demonstrate that the representation transfer performance drops significantly without the predictor. These findings emphasize the significant impact of cross-modal representation transfer on enhancing overall performance improvement and highlight the crucial role of BYOL's structure configuration in learning skeleton representations.

TABLE X
COMPARISON OF DIFFERENT FRAMEWORK DESIGNS USING ONE-SHOT EVALUATION PROTOCOL.

Framework	NTU-120	NTU-60	PKU-MMD
Without moving average target	68.0	80.2	84.3
Without predictor	63.8	76.6	81.1
Default	**70.4**	**83.9**	**87.3**

TABLE XI
COMPARISON OF DIFFERENT CLASSIFIERS USING ONE-SHOT EVALUATION PROTOCOL

Classifier	NTU-120	NTU-60	PKU-MMD
NN	66.7	80.2	84.5
LC	68.1	81.4	85.6
CC	**70.4**	**83.9**	**87.3**

Different Classifiers. We perform ablation experiments with different classifiers. As shown in Table XII, "NN" represents the nearest neighbor classifier, and "LC" represents the linear classifier, while "CC" stands for the cosine classifier. In the nearest neighbor classifier, the Euclidean distance is employed to measure the similarity between feature vectors and classify them. However, as the feature extraction of the skeleton graph using graph convolution occurs in a non-Euclidean space, it may not be optimal to utilize the Euclidean distance, as evidenced by its effect on the accuracy rate. The cosine classifier, on the other hand, employs cosine similarity. With limited support set samples, reducing the intra-class variance results in an improvement in accuracy compared to the linear classifier.

Different Video-Text Models. In addition to X-CLIP, we conduct experiments using different video-text models in the representation transfer framework to explore their impact. ActionCLIP [62], based on CLIP, employs a retrieval-based approach for video action recognition. Florence [63] further extends the CLIP method by leveraging a unified contrastive objective. As shown in Table XII, our method yields significant performance gains for the skeleton encoder by leveraging large-scale vision-language models, with the influence of more powerful video-text models becoming increasingly apparent.

Different Hyperparameters. We investigate the impact of hyperparameter modifications on the representation transfer framework using the one-shot evaluation protocol on the NTU-120 dataset. The vision-language embedding serves as the predictive target for the skeleton network during transfer, with its weight being determined as an exponential moving average of the skeleton projector weights. As shown in Figure 6, we can find that when the target decay rate is set to 1, the vision-language projector remains fixed to its initial random weights, leading to stable training but reducing the efficacy of pre-training. On the other hand, a target decay rate of 0 results in the immediate updating of the vision-language projector weights, causing instability during pre-training and a subsequent decline in the model performance. Intermediate values of the target decay rate (e.g., 0.5 and 0.999) also produce unsatisfactory results. It is worth noting that the proposed method achieves the best performance when λ is set to 0.99. Thus we set $\lambda = 0.99$ in the main experiment in our paper. The results further demonstrate that supervision from the vision-

TABLE XII
COMPARISON OF DIFFERENT VIDEO-TEXT MODELS USING ONE-SHOT EVALUATION PROTOCOL

Model	NTU-120	NTU-60	PKU-MMD
X-CLIP	70.4	83.9	87.3
ActionCLIP [62]	70.8	84.0	87.6
X-Florence [52]	**71.6**	**84.4**	**88.3**

language pre-trained model enhances the performance of the skeleton encoder, with the greatest improvement observed at the BYOL default value.

Fig. 6. Accuracy variation with different target decay rate

VI. CONCLUSION

This paper presents a novel cross-modal contrastive pre-training approach for few-shot skeleton action recognition. To address the challenge of pre-training stemming from the scarcity of large-scale skeleton datasets, we propose the representation transfer framework, which leverages the well-embedding text and video features from the vision-language pre-trained models to guide the learning of the skeleton encoder. Extensive experiments over different datasets demonstrate that our proposed method achieves superior few-shot skeleton action recognition performance.

REFERENCES

[1] S. Miao, Y. Hou, Z. Gao, M. Xu, and W. Li, "A central difference graph convolutional operator for skeleton-based action recognition," *TCSVT*, vol. 32, no. 7, pp. 4893–4899, 2021.
[2] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *TCSVT*, vol. 31, no. 11, pp. 4394–4408, 2021.
[3] 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 *ICCV*, pp. 13359–13368, 2021.
[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, vol. 32, 2018.
[5] 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.
[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, pp. 4077–4087, 2017.
[7] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.
[8] R. Memmesheimer, S. Häring, N. Theisen, and D. Paulus, "Skeleton-dml: deep metric learning for skeleton-based one-shot action recognition," in *WACV*, pp. 3702–3710, 2022.
[9] L. Wang and P. Koniusz, "Temporal-viewpoint transportation plan for skeletal few-shot action recognition," in *ACCV*, pp. 4176–4193, 2022.
[10] N. Ma, H. Zhang, X. Li, S. Zhou, Z. Zhang, J. Wen, H. Li, J. Gu, and J. Bu, "Learning spatial-preserved skeleton representations for few-shot action recognition," in *ECCV*, pp. 174–191, 2022.
[11] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *ICML*, pp. 894–903, PMLR, 2017.
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, 2009.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, pp. 4171–4186, 2019.
[14] 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.
[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, pp. 1597–1607, 2020.
[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, pp. 9729–9738, 2020.
[17] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
[18] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," in *NeurIPS*, p. 22243–22255, 2020.
[19] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, "Seed: Self-supervised distillation for visual representation," in *ICLR*, 2021.
[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," in *NIPS*, vol. 33, pp. 21271–21284, 2020.
[21] Z. Shao, Y. Li, and H. Zhang, "Learning representations from skeletal self-similarities for cross-view action recognition," *TCSVT*, vol. 31, no. 1, pp. 160–174, 2020.
[22] C. Wu, X.-J. Wu, T. Xu, Z. Shen, and J. Kittler, "Motion complement and temporal multifocusing for skeleton-based action recognition," *TCSVT*, 2023.
[23] 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.
[24] 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.
[25] 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*, vol. 36, pp. 762–770, 2022.
[26] O. Moliner, S. Huang, and K. Åström, "Bootstrapped representation learning for skeleton-based action recognition," in *CVPR*, pp. 4154–4164, 2022.
[27] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," in *NIPS*, vol. 29, 2016.
[28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, pp. 1126–1135, 2017.
[29] Y. Lu, L. Wen, J. Liu, Y. Liu, and X. Tian, "Self-supervision can be a good few-shot learner," in *ECCV*, pp. 740–758, 2022.
[30] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *PR*, vol. 130, p. 108792, 2022.
[31] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, pp. 8808–8817, 2020.
[32] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *CVPR*, pp. 4136–4145, 2020.
[33] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *TCSVT*, vol. 31, no. 3, pp. 1091–1102, 2020.
[34] L. Zhang, L. Zuo, Y. Du, and X. Zhen, "Learning to adapt with memory for probabilistic few-shot learning," *TCSVT*, vol. 31, no. 11, pp. 4283–4292, 2021.
[35] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *ICLR*, May 2019.
[36] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *ICLR*, April 2019.
[37] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *ICCV*, pp. 9062–9071, 2021.

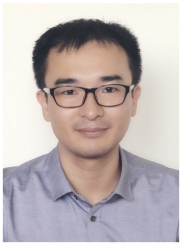
[38] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: a good embedding is all you need?," in *ECCV*, pp. 266–282, 2020.
[39] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *CVPR*, pp. 10618–10627, 2020.
[40] S. Liu, M. Jiang, and J. Kong, "Multidimensional prototype refactor enhanced network for few-shot action recognition," *TCSVT*, vol. 32, no. 10, pp. 6955–6966, 2022.
[41] X. Wang, W. Ye, Z. Qi, G. Wang, J. Wu, Y. Shan, X. Qie, and H. Wang, "Task-aware dual-representation network for few-shot action recognition," *TCSVT*, 2023.
[42] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational crosstransformers for few-shot action recognition," in *CVPR*, pp. 475–484, 2021.
[43] Y. Huang, L. Yang, and Y. Sato, "Compound prototype matching for few-shot action recognition," in *ECCV*, pp. 351–368, 2022.
[44] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *ECCV*, pp. 525–542, 2020.
[45] S. Kumar Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain, "Protogan: Towards few shot learning for action recognition," in *ICCV*, pp. 0–0, 2019.
[46] L. Zhu and Y. Yang, "Label independent memory for semi-supervised few-shot video classification," *TPAMI*, vol. 44, no. 1, pp. 273–285, 2020.
[47] L. Zhu and Y. Yang, "Compound memory networks for few-shot video classification," in *ECCV*, pp. 751–766, 2018.
[48] 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.
[49] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A. C. Murillo, "One-shot action recognition in challenging therapy scenarios," in *CVPR*, pp. 2777–2785, 2021.
[50] R. Memmesheimer, N. Theisen, and D. Paulus, "Sl-dml: Signal level deep metric learning for multimodal one-shot action recognition," in *ICPR*, pp. 4573–4580, 2021.
[51] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, and W. Zhang, "Smam: Self and mutual adaptive matching for skeleton-based few-shot action recognition," *TIP*, vol. 32, pp. 392–402, 2022.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, pp. 716–723, 2013.
[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, May 2020.
[58] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *CVPR*, pp. 1647–1656, 2017.
[59] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *TIP*, vol. 27, no. 4, pp. 1586–1599, 2017.
[60] Z. Wang, M. Fredrikson, and A. Datta, "Robust models are more interpretable because attributions look normal," in *ICML*, vol. 162, pp. 22625–22651, 2021.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
[62] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.
[63] 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.

VII. Biography Section

Mingqi Lu received the B.S. degree in the school of automation from Southeast University, Nanjing, Jiangsu Province, China, in 2018. Now, she is currently working towards the P.h.D. degree with the School of Automation, Southeast University, Nanjing, Jiangsu Province, China. Her current research interests include computer vision and action recognition.

Siyuan Yang received the BEng degree from Harbin Institute of Technology and the MSc degree from Nanyang Technological University. He is currently pursuing the Ph.D. degree with the Interdisciplinary Graduate Programme, Nanyang Technological University. His research interests include computer vision, action recognition, and human pose estimation.

Xiaobo Lu is a professor at the School of Automation and the deputy director of the Detection Technology and Automation Research Institute in Southeast University. He is a coauthor of the book An Introduction to the Intelligent Transportation Systems (Beijing: China Communications Press, 2008). He has earned many research awards, such as the first prize in Natural Science Award of the Ministry of Education of China and the prize in Science and Technology Award of Jiangsu province. His research interests include image processing, signal processing, pattern recognition, and computer vision.



Jun Liu (Member,IEEE) is an Assistant Professor with Singapore University of Technology and Design and an adjunct Assoicate Professor of the University of Western Australia. His research interests include computer vision and artificial intelligence. He has served as an Area Chair of NeurIPS, ICML, ICLR and CVPR. He is an Associate Editor of IEEE Transactions on Image Processing and IEEE Transactions on Biometrics, Behavior, and Identity Science.