Informative Sample Selection Model for Skeleton-Based Action Recognition With Limited Training Samples

Zhigang Tu, *Senior Member, IEEE*, Zhengbo Zhang, *Member, IEEE*, Jia Gong, Junsong Yuan, *Fellow, IEEE*, and Bo Du, *Senior Member, IEEE*

***Abstract*—Skeleton-based human action recognition aims to classify human skeletal sequences, which are spatiotemporal representations of actions, into predefined categories. To reduce the reliance on costly annotations of skeletal sequences while maintaining competitive recognition accuracy, the task of 3D Action Recognition with Limited Training Samples, also known as semi-supervised 3D Action Recognition, has been proposed. In addition, active learning, which aims to proactively select the most informative unlabeled samples for annotation, has been explored in semi-supervised 3D Action Recognition for training sample selection. Specifically, researchers adopt an encoder–decoder framework to embed skeleton sequences into a latent space, where clustering information, combined with a margin-based selection strategy using a multi-head mechanism, is utilized to identify the most informative sequences in the unlabeled set for annotation. However, the most representative skeleton sequences may not necessarily be the most informative for the action recognizer, as the model may have already acquired similar knowledge from previously seen skeleton samples. To solve it, we reformulate Semi-supervised 3D action recognition via active learning from a novel perspective by casting it as a Markov Decision Process (MDP). Built upon the MDP framework and its training paradigm, we train an informative sample selection model to intelligently guide the selection of skeleton sequences for annotation. To enhance the representational capacity of the factors in the state-action pairs within our method, we project them from Euclidean space to hyperbolic space. Furthermore, we introduce a meta tuning strategy to accelerate the deployment of our method in real-world scenarios. Extensive experiments on three 3D action recognition benchmarks demonstrate the effectiveness of our method.**

***Index Terms*—Human action recognition, video analysis.**

I. Introduction

HUMAN action recognition [1], [2], [3] aims to classify actions from spatiotemporal sequences into a set of predefined categories, and serves as a core component in various applications such as activity understanding [4], interaction modeling [5], and robotic control [6]. Based on the type of input data, existing action recognition methods can be broadly categorized into RGB-based [7], [8], [9], depth-based [10], and 3D skeleton-based approaches [11], [12], [13]. Among these data, 3D skeleton data, which represents the human body as a set of keypoints in 3D space, has received increasing attention in recent years [5], [14]. This is because, compared to RGB or depth data, skeleton representations offer compact, high-level descriptions of human motion and are generally more robust to appearance variations, background clutter, and viewpoint changes [15]. Furthermore, skeleton sequences can be efficiently captured by commodity depth sensors, enabling the development of numerous supervised methods to learn spatiotemporal features for skeleton-based action recognition [16].

Recently, with the rapid development of deep learning [9], [19], [20], [21], deep neural networks [9], [22], [23], [24], [25], [26] have been extensively explored for modeling spatiotemporal representations of skeleton sequences in supervised settings. Specifically, Recurrent Neural Networks have been applied to capture temporal dependencies in action sequences [20], while Convolutional Neural Networks have been adopted by transforming joint coordinates into 2D maps [1]. Besides, Graph Convolutional Networks have gained popularity due to their strong performance in modeling the structured topology of skeleton data [2]. Despite their promising results, these methods typically rely on large-scale annotated datasets for training [20]. However, obtaining accurate frame-level annotations is often labor-intensive and requires annotators to possess expert knowledge of human skeletons, which poses significant challenges to scalability and generalization across unseen actions or subjects [17].

To reduce the reliance on extensive annotations without sacrificing recognition accuracy in skeleton-based action

Received 6 August 2025; accepted 27 October 2025. Date of publication 7 November 2025; date of current version 13 November 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3015600 and in part by the Fundamental Research Funds for Central Universities under Grant 2042023KF0180 and Grant 2042025KF0053. The associate editor coordinating the review of this article and approving it for publication was Dr. Jie Wen. *(Corresponding author: Zhengbo Zhang.)*

Zhigang Tu is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: tuzhigang@whu.edu.cn).

Zhengbo Zhang is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: zhangzb@whu.edu.cn).

Jia Gong is with Shanghai Academy of AI for Science, Shanghai 200232, China.

Junsong Yuan is with the Department of Computer Science and Engineering, University at Buffalo, State University of New York, Buffalo, NY 14228 USA.

Bo Du is with the School of Computer Science, Wuhan University, Wuhan 430072, China.

Digital Object Identifier 10.1109/TIP.2025.3627418

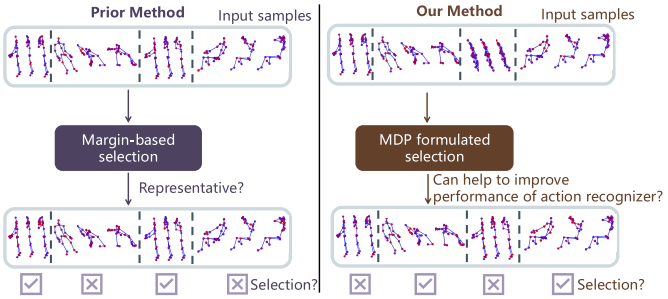

Fig. 1. Motivation of our method. Previous semi-supervised 3D action recognition via active learning (S3ARAL) approaches [17] rely on margin-based selection strategy that aims to identify representative samples. However, such strategy may select samples with limited novel information, leading to sub-optimal performance of the trained action recognizer. In contrast, we propose a novel perspective by reformulating S3ARAL as a Markov Decision Process (MDP) [18]. Within this framework, we train an informative sample selection model to intelligently choose training samples that are more likely to improve the action recognizer's performance, thereby enabling the training of a more effective model.

recognition, semi-supervised learning approaches have been proposed, and received considerable attention. Specifically, the semi-supervised setting assumes that annotated auxiliary classes are unavailable. Methods such as ASSL [27], MS^{2}L [28], and SC3D [29] aim to learn informative representations from both labeled and unlabeled data while preserving classification performance. Nevertheless, these methods typically overlook the fact that not all labeled samples contribute equally to model training. Selecting the representative samples for annotation can enhance classifier effectiveness while further reducing annotation cost [17].

To address this issue, the previous work explores AL-SAR by incorporating active learning (AL), which aims to proactively select the most informative unlabeled samples for annotation, with semi-supervised 3D action recognition [17]. Specifically, AL-SAR first employs an encoder–decoder framework to encode skeleton sequences into a latent space. It then leverages the clustering information in this latent space, together with a margin-based selection strategy built on a multi-head mechanism, to identify informative skeleton sequences in the unlabeled set for annotation. However, the representative skeleton sequences may not always be the most informative for the action recognizer, as the model may have already learned similar patterns from previously labeled samples. Consequently, selecting such redundant samples may offer limited benefit and lead to sub-optimal performance for the action recognizer. Hence, it is crucial to boost Semi-supervised 3D Action Recognition via AL (S3ARAL) in a more intelligent and effective manner.

In this paper, inspired by [30], we handle this problem from *a novel perspective* by formulating S3ARAL as a Markov Decision Process (MDP) [18] (see Figure 1). Within this MDP framework, the action corresponds to selecting the representative samples for annotation. Our key insight of such formulation is that MDP and S3ARAL share a common objective of maximizing long-term benefits: MDP aims to maximize expected cumulative rewards, while S3ARAL seeks to maximize action recognizer's performance by selecting the informative samples for annotation. Moreover, MDP offers solid theoretical foundations [18] and has demonstrated its capability to make informed decisions (action) in complex state spaces across multiple domains, like robotics [31] and Game AI [32]. Based on our formulation, we set the reward in the MDP as the action recognizer's performance improvement among training stages, with the objective of maximizing cumulative performance gain of the recognizer.

In addition, human skeleton data naturally form tree-like graphs, where hierarchical relationships exist among joints. As a result, directly computing features of the skeleton data in the Euclidean space may fail to capture these structural properties effectively. Motivated by the exponential volume growth of hyperbolic space [33], which enables efficient representation of hierarchical skeleton data, we project the factors in the state-action space of our MDP formulation from the Euclidean space to the hyperbolic space to enhance their representational capacity. Furthermore, when our method is deployed in real-world scenarios, it typically requires re-training on the enlarged labeled dataset, which can be time-consuming. To address this issue, we design a meta-tuning strategy based on meta learning [34] to improve the generalizability of the proposed method, allowing a rapid deployment.

To ensure a fair comparison with the previous S3ARAL method [17], we follow its experimental setup and conduct experiments on three commonly used 3D action recognition datasets: UWA3D Multiview Activity II [35], North-Western UCLA [36], and NTU RGB+D 60 [14]. Extensive experiments on these datasets validate the effectiveness of our method.

The main contributions are summarized as follows:

- Existing S3ARAL method often overlooks whether the selected samples are informative for the action recognizer, leading to sub-optimal performance of the trained model. To address this issue, we approach the task from a new perspective by formulating it as a MDP, and design a novel training framework that directly links the improvement of the recognizer's performance with the informativeness of the selected samples, enabling the training of a more powerful action recognizer.
- To enhance the representational capacity of factors in the state-action pairs of our MDP framework, we map them from the Euclidean space to the hyperbolic space. Besides, we introduce a meta tuning strategy to accelerate the deployment of our method in real-world scenarios.
- We conduct extensive experiments on three 3D action recognition datasets, where the results show that our method effectively selects informative samples, enabling the training of a high-performing action recognizer. Moreover, our method exhibits strong generalization ability and consistently achieves state-of-the-art performance across all three datasets.

II. Related Work

A. Human Action Recognition

Existing human action recognition methods [3], [8], [12], [13], [37] can be broadly categorized into two groups: RGB-based methods [8] and skeleton-based methods [3], [12],

[13]. Due to the compact yet informative nature of skeleton sequences for representing human behavior, skeleton-based action recognition has attracted substantial research interest.

Early skeleton-based action recognition methods explore a variety of network architectures to process skeleton sequences. For instance, Liu et al. [5] introduce a spatial-temporal LSTM for skeleton-based action recognition, while Ke et al. [19] propose converting skeleton sequences into grayscale images, which are then fed into a CNN, benefiting from the advances in deep learning [38], [39], [40], [41], [42]. Over time, graph convolutional networks (GCNs) have emerged as a dominant architecture for this task. Yan et al. [39] pioneer this direction with ST-GCN, the first GCN-based framework for skeleton-based action recognition. Building upon this, various enhanced GCN models have been proposed, including AS-GCN [40], and CTR-GCN [43]. More recently, transformer-based models have gained popularity for this task. Several works [41], [44] have proposed end-to-end transformer architectures for skeleton-based action recognition, such as the 3Mformer [41] and the UPS [44].

Different from these methods, this work tackles a relatively new task: Semi-supervised 3D Action Recognition via Active Learning. It selects informative human sequence data from the unlabeled set through active learning and trains an action recognizer based on the annotated samples.

B. Semi-Supervised Human Action Recognition

Semi-supervised learning for action recognition aims to extract motion representations from unlabeled data by regularizing features or solving pretext tasks. Many studies have proposed semi-supervised methods tailored for unlabeled RGB data [4], [45], [46]. Specifically, [45] focus on predicting motion flows to learn video representations, while [4] introduces a convolutional auto-encoder that separately captures spatial and temporal information from raw videos. Leveraging the observation that varying frame rates do not alter the semantic meaning of actions, [46] designs a two-stream contrastive model in the temporal domain to utilize unlabeled videos at different playback speeds.

Although these methods achieve promising results on RGB video data, they are less suitable for long-term skeleton sequences. Unlike RGB videos, which contain rich appearance and scene context, skeleton data primarily encodes human motion. Thus, effectively capturing joint and bone movement patterns is critical for semi-supervised skeleton action recognition. To this end, Zheng et al. [47] proposes a joint inpainting approach to learn features from unlabeled skeleton sequences, followed by the work of Si et al. [48]. However, such self-supervised methods, which rely on inpainting key joints, fail to capture holistic motion dynamics and are not well-suited for GCN-based architectures. Lin et al. [28] introduce a multi-task self-supervised framework for skeleton data, but their model struggled to extract joint-bone fused features using a network structure optimized for action recognition. Moreover, [28], [47], and [48] still adopt RNN-based backbones, which are suboptimal for capturing spatial-temporal dependencies.

In this paper, unlike previous semi-supervised skeleton-based action recognition methods that often randomly select a batch of samples for annotation to train the action recognizer, we adopt active learning to select informative skeleton samples for annotation.

C. Markov Decision Process

Markov Decision Process (MDP) has become the standard model for studying how agents make optimal decisions in uncertain environments [18], [31], [32], [49]. In recent years, the emergence of deep reinforcement learning (DRL) [31] has made solving complex high-dimensional MDP problems possible. MDP based on DRL has been widely applied to numerous practical problems including robotic control [32], [49], medical decision-making [18], etc, demonstrating powerful modeling capabilities and potential for solving complex decision problems. In this paper, we are the first to formulate semi-supervised 3D action recognition via active learning as a Markov Decision Process (MDP), and we train an informative sample selection model within the MDP framework to intelligently select samples for annotation.

III. Method

Given an unlabeled set of skeleton sequences and a limited annotation budget, the goal of the Semi-supervised 3D Action Recognition via AL (S3ARAL) task is to iteratively select the informative samples for annotation to maximize the performance of the target action recognizer.

Motivated by the similarity in objectives between S3ARAL and the Markov Decision Process (MDP)—where both aim to maximize long-term returns, with S3ARAL focusing on improving the action recognizer's performance throughout the active learning process and MDP targeting the maximization of expected cumulative rewards—we propose a novel perspective by formulating S3ARAL as an MDP (Section III-B). Specifically, we design a informative sample selection model that selects informative yet diverse skeleton sequences for annotation (Section III-C). Moreover, to enable quick adaptation to the enlarged labeled skeleton set during deployment, we further introduce a meta tuning strategy based on meta-learning [34] (Section III-D). We introduce the overall training and testing of our method in Section III-E. The pipeline of our method is illustrated in Figure 2.

A. Task Settings and Notation

To improve clarity, in this subsection we present the task setting and introduce the notations of the key symbols used in our method.

1) Task Settings: In this paper, we address the task of Semi-supervised 3D Action Recognition via Active Learning (S3ARAL). S3ARAL operates on datasets composed of multi-dimensional time series that capture the coordinates of body keypoints over time. We denote the dataset as $X = \{X^U \cup X^L\}$, where X^U and X^L represent the unlabeled and labeled subsets, respectively. Each sample $x_i \in X$ is a sequence $x_i = [x_1, x_2, \ldots, x_T]$, where $x_t \in \mathbb{R}^{p \times d}$ denotes the coordinates of p keypoints at time t, and d is the dimensionality of the keypoints (typically $d = 3$). The number of keypoints p may vary across different datasets.

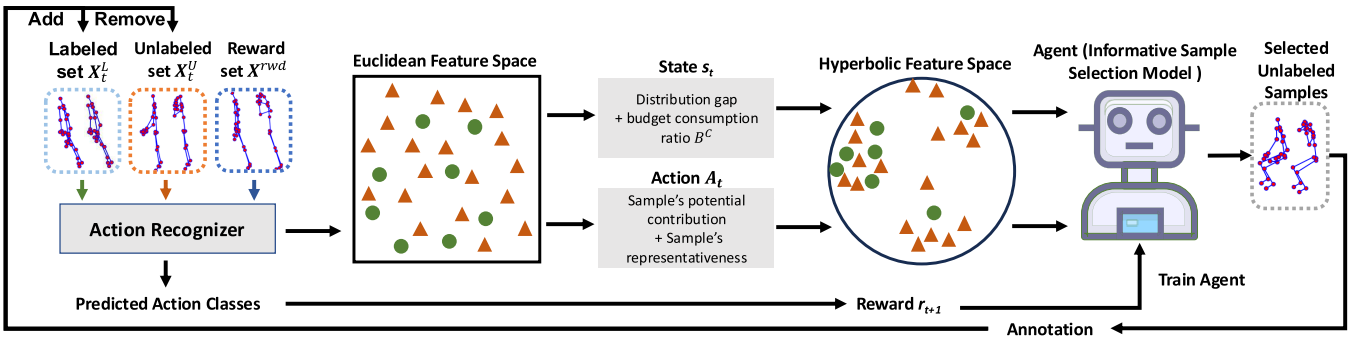

Fig. 2. The overall pipeline of our method. In the task of semi-supervised 3D action recognition via active learning, given an unlabeled dataset and an annotation budget, we divide the dataset into a labeled set, an unlabeled set, and a reward set. We then design an Informative Sample Selection Model (ISSM) to select informative samples for annotation. The annotated samples are used to train the action recognizer. To ensure that the ISSM receives sufficient information for effective sample selection, we carefully construct the state-action pairs. The state encodes the distribution gap between the labeled and unlabeled sets, along with the budget consumption ratio. The action captures both the sample's potential contribution to improving the action recognizer and its representativeness. To enhance the expressiveness of the state-action representations for the skeleton-based action recognition task, we project them from the Euclidean space to hyperbolic space. This is motivated by the exponential volume growth of hyperbolic space, which is well-suited for modeling the hierarchical structure of human skeletons. The performance improvement of the action recognizer between consecutive iterations is treated as the reward for training the ISSM.

Specifically, in the S3ARAL task, all samples are initially unlabeled, i.e., $X = X^U$. With an annotation budget B, at the t-th iteration, given an unlabeled set X_t^U, a labeled set X_t^L, and an action recognizer AR_t, the S3ARAL method proceeds as follows: (1) Assess the informativeness of the unlabeled samples; (2) Select a batch of informative skeleton sequences for annotation; (3) Transfer the selected samples from X_t^U to X_t^L, and retrain the action recognizer AR_t on the updated labeled set X_{t+1}^L to obtain a new model AR_{t+1}.

2) Notation: Next, we introduce the meanings of some key notations used in our paper.

Skeleton Sequence. $x = [x_1, x_2, \ldots, x_T]$, where T is the sequence length and $x_t \in \mathbb{R}^{p \times d}$ denotes the 3D coordinates of p joints at time step t, with $d = 3$.

Dataset. $X = \{X^U \cup X^L\}$, composed of labeled and unlabeled subsets.

Budget. B, the maximum number of skeleton sequences that can be annotated.

State. $[\widetilde{\text{MMD}}_t(S^L, S^U), B_t^C]$ where $\widetilde{\text{MMD}}_t$ is the hyperbolic projection of the distribution gap and B_t^C is the budget consumption ratio.

Action. a_t, selection of a sample based on its state.

Reward. r_{t+1}, the accuracy gain of the action recognizer between iterations, measured on a reward set X^{rwd}.

Transition. $(s_t, a_t, r_{t+1}, s_{t+1})$, which updates the labeled and unlabeled sets.

B. Semi-Supervised 3D Action Recognition via Active Learning as a Markov Decision Process

To solve the challenge Semi-supervised 3D Action Recognition via Active Learning (S3ARAL) task, previous S3ARAL method [17] encodes skeleton sequences into a high-dimensional latent space and group the resulting latent representations into clusters. Informative samples are then selected for annotation based on their corresponding uncertainty scores and their distances within the latent space clusters in the feature space. However, representative skeleton sequences may not always be the most informative for the action recognizer, as the model may have already learned similar patterns from previously labeled samples. Hence, selecting such redundant samples may provide limited benefit and result in sub-optimal recognition performance. To address this, inspired by [30], we approach S3ARAL from a novel perspective by formulating it as a Markov Decision Process (MDP), motivated by their shared objective of maximizing future returns. Besides, we treat the performance improvement of the action recognizer as the reward in our MDP. Such formulation enables a more intelligent sequence sample selection strategy.

Below, we first provide an explanation of how to view the S3ARAL task from the MDP perspective.

1) New Perspective on S3ARAL Task: In this paper, we approach S3ARAL from the novel MDP perspective to enable more intelligent sample selection for annotation. Specifically, the S3ARAL process is formulated as a MDP tuple $(s_t, a_t, r_{t+1}, s_{t+1})$, with the core steps reinterpreted as follows: (1) State Estimation: Estimate the state s_t, which captures the distributional discrepancy between the unlabeled set X_t^U and the labeled set X_t^L at iteration t. (2) Action Selection: Evaluate each state-action pair (s_t, a_t) to identify the informative skeleton sequence for annotation. (3) Environment Transition: Update X_t^L and X_t^U to X_{t+1}^L and X_{t+1}^U by moving the selected skeleton sequence from X_t^U to X_t^L. Retrain the action recognizer AR_t on the updated labeled set X_{t+1}^L to obtain a new model AR_{t+1}, and update the state to s_{t+1}. (4) Reward Computation: Compute the reward r_{t+1} based on the performance improvement of AR_{t+1} over AR_t, evaluated on a separate reward set X^{rwd}, to guide the agent's update.

In this way, we reformulate S3ARAL task as a MDP, allowing us to leverage the MDP framework with solid theoretical foundations [31] to perform the S3ARAL task in a more intelligent manner.

C. Informative Sample Selection Model for Semi-Supervised 3D Action Recognition via Active Learning

In the previous section, we introduce how the Semi-supervised 3D Action Recognition via Active Learning (S3ARAL) task can be viewed from the perspective of the

Markov Decision Process (MDP). In this section, we present the details of our proposed informative sample selection model, which enables us to perform the S3ARAL task from the MDP perspective.

In this work, we adopt the widely used Q-learning algorithm [31] as our MDP framework, where the designed informative sample selection model (ISSM), i.e., the agent in the MDP, performs batch sampling to select skeleton samples for annotation. Specifically, our ISSM evaluates each state-action pair (s_t, a_t) and selects the action a_t with the highest associated Q-value. By deriving the reward directly from the improvement of the action recognizer, ISSM is trained to learn a policy that maximizes both the cumulative reward and the performance of the action recognizer. Besides, to support ISSM in making informed decisions, we design a state-action representation based on hyperbolic space [50]. This hyperbolic state design provides exponential volume growth, making it well-suited for representing the tree-like structure of human skeletons in a compact and expressive manner [51].

Below, we sequentially introduce the design of the state, action, and reward, along with the training strategy within our MDP framework.

1) State: Intuitively, the state s_t should capture the distribution gap between the labeled dataset X_t^L and the unlabeled dataset X_t^U, enabling the agent to identify the informative skeleton sequence that can mitigate the distribution shift between X_t^L and X_t^U. The purpose of computing this distribution gap is to make the training set more unbiased, thereby increasing the likelihood that the trained action recognizer generalizes well to unseen cases. Besides, skeleton data naturally form tree-like graphs, where hierarchical relationships exist among joints [14]. Therefore, computing features in Euclidean space may fail to capture these structural properties effectively. In contrast, hyperbolic space exhibits exponential volume growth, allowing for efficient representation of hierarchical skeleton data. Hence, to enhance the representation of the factors included in our state and action space, we transform the computed distribution gap from the Euclidean space to hyperbolic space.

Below, we first introduce how to compute the distribution gap between the labeled and unlabeled datasets, followed by the definition of hyperbolic space and the projection of the computed distribution gap from the Euclidean space to hyperbolic space.

i) Distribution Gap. To model the distributional drift between the labeled dataset X_t^L and the unlabeled dataset X_t^U, we treat them as two separate domains and measure the domain gap accordingly. Specifically, we adopt the Maximum Mean Discrepancy (MMD) to quantify the discrepancy, as follows:

$$\begin{aligned}\mathrm{MMD}(S^L, S^U) = &\sum_{i=1}^{n_L}\sum_{j=1}^{n_L}\frac{k(p_i,p_j)}{n_L^2} + \sum_{i=1}^{n_U}\sum_{j=1}^{n_U}\frac{k(q_i,q_j)}{n_U^2} \\ &- \sum_{i=1}^{n_L}\sum_{j=1}^{n_U}\frac{2\cdot k(p_i,q_j)}{n_L n_U},\end{aligned} \tag{1}$$

where S^L and S^U denote the feature space distributions of X_t^L and X_t^U, respectively, and the computed $\mathrm{MMD}(S^L, S^U)$ is a scalar representing the discrepancy between them. The samples from S^L and S^U are denoted as p and q, respectively. n_L and n_U are the numbers of samples in X_t^L and X_t^U. $k(\cdot)$ is a radial basis kernel used to compute pairwise distances.

ii) Hyperbolic Space. Here, inspired by [33], we adopt the Poincaré model [50] as the hyperbolic space. Specifically, an n-dimensional Poincaré model, denoted as $\mathbb{B}_\kappa^n$, is a Riemannian manifold $(\mathbb{B}_\kappa^n, g_{\mathbf{x}}^{\mathbb{B}})$ with constant negative curvature $\kappa < 0$. The Poincaré model is defined as:

$$\mathbb{B}_\kappa^n = \left\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < -\frac{1}{\kappa}\right\}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm. The manifold is further equipped with the Riemannian metric tensor $g_{\mathbf{x}}^{\mathbb{B}}$:

$$g_{\mathbf{x}}^{\mathbb{B}} = \left(\frac{2}{1+\kappa\|\mathbf{x}\|^2}\right)^2 g^{\mathbb{E}}, \tag{3}$$

where $\mathbf{x} \in \mathbb{B}_\kappa^n$ and $g^{\mathbb{E}}$ denotes the Euclidean metric tensor. This formulation highlights the strength of hyperbolic space in modeling hierarchical and structured data, where distances grow exponentially, similar to a tree structure. Such a property makes it particularly suitable for skeleton-based human action recognition, as human motion inherently follows a multi-scale hierarchical pattern.

iii) Projection to Hyperbolic Space. After introducing definition of the hyperbolic space, which is more suitable for modeling human skeletons, we then present how to project the distribution gap computed in Euclidean space into the hyperbolic space. Specifically, given the distribution gap $\mathrm{MMD}(S^L, S^U)$ computed in Euclidean space, we follow [50] to map it into hyperbolic space through the exponential map $\exp_{\mathbf{0}}^c(\cdot)$ as:

$$\begin{aligned}&\widetilde{\mathrm{MMD}}(S^L, S^U) \\ &= \exp_{\mathbf{0}}^c\left(\mathrm{MMD}(S^L, S^U)\right) \\ &= tanh(\sqrt{c}\|\mathrm{MMD}(S^L, S^U)\|)\frac{\mathrm{MMD}(S^L, S^U)}{\sqrt{c}\|\mathrm{MMD}(S^L, S^U)\|}.\end{aligned} \tag{4}$$

Here, $\widetilde{\mathrm{MMD}}(S^L, S^U)$ in the above equation denotes the projected distribution gap $\mathrm{MMD}(S^L, S^U)$.

Moreover, the available budget in our MDP framework is a critical factor for our ISSM to make effective selections. To capture this, we incorporate the budget consumption ratio B^C, which is an indicator of the remaining annotation resources, into the state representation. Hence, our state s_t is defined as $[\widetilde{\mathrm{MMD}}_t(S^L, S^U), B_t^C]$, capturing both the distribution gap between the labeled and unlabeled sets and the current budget status. This state representation guides the ISSM in selecting the types of human skeletons that are beneficial for improving the performance of the action recognizer.

2) Action: The action should ideally capture the potential contribution of a specific unlabeled human skeleton when it is added to the labeled set X_t^L. Intuitively, by combining the state and action representations, our ISSM is expected to have sufficient information to evaluate each unlabeled human skeleton and select the informative skeleton from the unlabeled set X_t^U for annotation. To this end, we associate each unique skeleton x from the unlabeled pool X_t^U with an action a_t in

the action space A_t. Specifically, to facilitate the selection of informative samples, we extract two types of features from each unlabeled skeleton sequence: (1) the skeleton sequence's potential contribution to improving the action recognizer, measured by its uncertainty. (2) the representativeness of the skeleton sequence within the unlabeled dataset. Below, we sequentially introduce the computation methods of these two types of action representations.

We adopt the widely used marginal index (MI) to measure the uncertainty [52]. Specifically, for a skeleton sample x, MI quantifies the confidence gap between the most and the second most probable action classes predicted by the action recognizer, and $\mathrm{MI}(x)$ is calculated as:

$$\mathrm{MI}(x) = 1 - \left(\max_{C} p(c \mid x) - \max_{C \neq c^*} p(c \mid x) \right), \tag{5}$$

where

$$c^* = \arg\max_{C} p(c \mid x). \tag{6}$$

Here, C denotes all action classes, c^* denotes the action class with the highest predicted probability by the action recognizer, and $\max_{c \neq c^*} p(c \mid x)$ denotes the second highest predicted probability.

We then present our parameterization of the skeleton sequence representativeness. Specifically, we select representative samples with respect to the unlabeled set X_t^U, by leveraging the distribution of similarity scores, and introduce a histogram-based representation R that encodes the cosine similarity distribution between a sample x and the average behavior of samples in X_t^U across the feature space of the action recognizer. Including such a factor in our action space allows the ISSM to avoid repeatedly selecting representative action sequences that have already been learned by the action recognizer, thereby improving sampling efficiency.

It is worth noting that for the features in the action space, we follow the approach used in the state representation and project them from Euclidean space to hyperbolic space.

3) Reward: The reward serves as a metric to quantify the benefit that a selected unlabeled skeleton sequence brings to our action recognizer AR_t at iteration t. To enable accurate reward estimation, we reserve a reward subset X^{rwd} prior to our active learning procedure. The reward r_{t+1} is defined as the accuracy gain of the action recognizer on X^{rwd}, computed as the difference in performance between AR_{t+1} and AR_t. Note that X^{rwd} is used solely for evaluation and is excluded from the training of the action recognizer. With the reward signal r_{t+1}, our ISSM can be optimized to select the informative skeleton sequence, thereby improving 3D action recognition accuracy in the active learning iteration.

4) Training of the ISSM: To train our ISSM, we adopt the Double DQN framework [31] and optimize the model by minimizing the temporal difference (TD) error:

$$\begin{aligned} TD(\theta, \hat{\theta}) = (&AR_t(s_t, a_t; \theta) - r_{t+1} \\ &- \gamma \cdot AR_{t+1}(s_{t+1}, a_{t+1}; \hat{\theta}))^2 \end{aligned} \tag{7}$$

Here, θ denotes the parameters of our ISSM, and $\hat{\theta}$ corresponds to the parameters of the target (off-policy) model, which maintains the learned Q-values and is periodically updated with θ, following the Double DQN setup.

Overall, the carefully designed state and action spaces of our ISSM, combined with the theoretically grounded MDP training framework, enable our ISSM to more intelligently select skeleton sequences for annotation.

D. Meta Tuning

When deploying our trained informative sample selection mode (ISSM) to real-world scenarios, it typically requires re-training on an enlarged labeled dataset. This re-training process could be time-consuming due to the continuously growing dataset size. To address this issue, we design a meta tuning strategy based on meta-learning [34], which aims to enhance the generalization ability of our ISSM and thereby accelerates its deployment. Specifically, our proposed meta tuning formulates the re-training of our ISSM on the expanded labeled dataset as a meta-learning task, which consists of three stages: virtual-train, virtual-test, and meta-update. We consider our ISSM training in source dataset, i.e., training on the unexpanded labeled dataset, as a virtual-train task, and re-trained our ISSM on the expanded labeled dataset as a virtual-test task. Virtual-test provides feedback on how to optimize the ISSM in a manner that generalizes across different datasets. Subsequently, the meta-update phase consolidates the feedback from the first two phases and applies the actual parameter updates to our model, resulting in improved generalization capabilities that promote fast adaptation.

To simulate the meta tuning process, we partition initial training dataset X_{i} into a smaller subset as the virtual-train set X_{i}^{vtr} and a larger subset as the virtual-test set X_{i}^{vte}. Our objective is to learn meta parameters θ_{mt} that can adapt effectively to X_{i}^{vte} after being optimized on X_{i}^{vtr}. Specifically, starting from randomly initialized meta parameters θ_{mt}, we first update our ISSM on X_{i}^{vtr} to obtain θ_{mt}^*. We then use the updated ISSM parameterized by θ_{mt}^* to perform active learning on X_{i}^{vtr} and compute the meta loss, which is used to refine θ_{mt}^*. The meta loss is the Temporal Difference (TD) error:

$$\begin{aligned} TD_{\mathrm{mt}}(\theta_{\mathrm{mt}}^*) = \sum_{t=0}^{H-1} (&AR_t(s_t, a_t; \theta_{\mathrm{mt}}^*) \\ &- r_{t+1} - \gamma \cdot AR_{t+1}(s_{t+1}, a_{t+1}; \theta_{\mathrm{mt}}^*))^2. \end{aligned} \tag{8}$$

Here, H denotes the number of iterations in our meta tuning process. After obtaining the meta loss $TD_{\mathrm{mt}}(\theta_{\mathrm{mt}}^*)$, we update the meta parameters according to:

$$\theta_{\mathrm{mt}} = \theta_{\mathrm{mt}} - \beta \cdot \nabla TD_{\mathrm{mt}}(\theta_{\mathrm{mt}}^*). \tag{9}$$

Here, β is the learning rate for meta tuning. By minimizing the meta loss TD_{mt}, we obtain the meta parameters θ_{mt}^* that enable fast adaptation to the virtual-test set X_{i}^{vte} using updates from only the virtual-train set X_{i}^{vtr}.

Overall, the meta tuning strategy enhances the generalization ability of our ISSM, thereby accelerating its deployment in real-world scenarios.

Algorithm 1 Overall Framework of the Proposed Method

Require: Unlabeled set X^U, budget B, initial labeled set X_0, reward set X^{rwd}
Ensure: Final trained action recognizer
Stage 1: Active Learning with Limited Budget
Initialize $X^L \leftarrow X_0$, $X^U \leftarrow X^U \setminus X_0$
while $|X^L| < B$ and $X^U \neq \varnothing$ **do**
    Construct state features
    Estimate quality scores of all candidates in X^U
    Select top samples and add to X^L
    Remove selected samples from X^U
    Retrain the recognition model on X^L
    Evaluate on X^{rwd} and update model
end while
Stage 2: Meta Tuning
Split X_0 into virtual training and validation subsets
for each meta-iteration **do**
    Update the model using virtual training set
    Validate on virtual validation set and refine parameters
end for
return Final trained recognition model

E. Overall Training and Testing

In the previous sections, we formulate Semi-supervised 3D Action Recognition via Active Learning (S3ARAL) as a Markov Decision Process (MDP), allowing us to leverage the well-established MDP framework to perform S3ARAL in a more intelligent manner (Section III-B). Then, we introduce the informative sample selection model (ISSM) for selecting skeleton data for annotation (Section III-C), as well as a meta-tuning strategy to enable rapid adaptation of ISSM in real-world scenarios (Section III-D). In this section, we provide an overview of the training and testing procedures of our method, along with an algorithm that outlines the overall framework (see Algorithm 1).

1) Training: Given an unlabeled dataset X and an annotation budget B, our method proceeds as follows: We first randomly sample an initial subset X_{i} for annotation. Using the labeled initial subset X_{i}, we then simulate the S3ARAL process by partitioning X_{i} and training our ISSM. Specifically, we divide the labeled initial set X_{i} into a labeled set X_{i}^L, an unlabeled set X_{i}^U, and a reward set X^{rwd}, and then allow our ISSM to select the skeleton sequence as described in Section III-C. Furthermore, once our ISSM is trained on X_{i}, it can be deployed to perform the actual S3ARAL process on the remaining unlabeled pool $X^U = X \backslash X_{\mathrm{i}}$, until the annotation budget B is exhausted. We refer to this stage as the deployment phase, during which our ISSM selects batch skeleton data $\{x^n\}_{n=1}^N$ from X^U for annotation at each iteration, gradually expanding the labeled pool $X^L = X^L \cup \{x^n\}_{n=1}^N$ to update our action recognizer AR. We initialize $X^L = X_{\mathrm{i}}$ at the beginning of this phase. Besides, to accelerate the deployment of our method in the deployment phase, we design a meta tuning strategy for our ISSM (Section III-D).

2) Testing: After deployment, our ISSM remains frozen and automatically selects samples for annotation.

IV. Experiments

A. Evaluation Dataset and Metric

For a fair comparison, we follow [17] to evaluate our method on three widely used benchmark datasets: UWA3D Multiview Activity II (UWA3D) [35], North-Western UCLA (NW-UCLA) [36], and NTU RGB+D 60 [14]. These datasets vary in the number of action classes and include both cross-view and cross-subject evaluation settings. We follow [17] and use classification accuracy as the evaluation metric. Below, we provide a detailed introduction to the three evaluation datasets used in this work.

UWA3D consists of 30 human action categories, with each action performed four times by ten subjects and recorded from four viewpoints: frontal, left, right, and top. Following [17], we use the frontal and left views for training and reserve the right view for testing to ensure a fair comparison. This setup presents a more challenging evaluation scenario due to the increased view discrepancy.

NW-UCLA is collected using Kinect v1 cameras and consists of 1,494 samples performed by 10 subjects. It includes 10 action classes, with each skeleton comprising 20 joints. Following the evaluation protocol in [17], we use samples from camera views 1 and 2 as the training set, while samples from camera view 3 serve as the test set.

NTU RGB+D 60, captured with a Kinect v2 sensor, is the largest publicly available benchmark for depth-based action recognition, containing over 56,000 video sequences and 4 million frames. It includes recordings from 80 different viewpoints and covers 60 action classes, ranging from daily activities and medical conditions to interactive actions, performed by 40 subjects aged between 10 and 35. The dataset poses considerable challenges due to large intra-class variation and diverse viewpoints. Owing to its scale, it is particularly suitable for deep learning-based activity recognition. Models pre-trained on this dataset can be effectively fine-tuned on smaller datasets, leading to faster convergence and improved performance. In the NTU RGB+D 60 dataset, each sample provides 3D coordinates of 25 body joints, offering rich skeletal information for detailed analysis.

B. Baseline Methods

To ensure a comprehensive and fair comparison with previous Semi-supervised 3D Action Recognition via Active Learning method, we follow the settings in [17] and compare our approach with state-of-the-art (SOTA) active learning and semi-supervised methods.

In particular, we compare our method against four representative active learning approaches: uniform sampling [17], core set selection [53], discriminator-based selection [54], and consistency-based active learning under augmentation [55]. Specifically, in uniform sampling, samples are randomly selected for annotation from the entire dataset. Core-set selection aims to cover the feature space by choosing samples that minimize the overall coverage radius. Discriminator-based selection employs a discriminator to distinguish whether a sample is labeled. Consistency-based active learning under

TABLE I

QUANTITATIVE COMPARISONS WITH REPRESENTATIVE ACTIVE LEARNING METHODS (UNIFORM SAMPLING [17], CORE SET SELECTION [53], DISCRIMINATOR-BASED SELECTION [54], AND CONSISTENCY-BASED ACTIVE LEARNING UNDER AUGMENTATION [55], AND AL-SAR [17]) ON THREE COMMON 3D ACTION RECOGNITION DATASETS (UWA3D [35], NW-UCLA [36], AND NTU RGB+D 60 [14]) UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES. "%LABELS" DENOTES THE PROPORTION OF LABELED SAMPLES RELATIVE TO THE TOTAL NUMBER OF TRAINING SAMPLES IN THE DATASET, WHILE "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. "VIEW3" REFERS TO USING THE THIRD CAMERA VIEW OF THE MULTI-VIEW UWA3D DATASET AS THE TESTING SET, WHILE THE REMAINING VIEWS ARE USED FOR TRAINING. "CS" STANDS FOR "CROSS-SUBJECT", MEANING THAT THE DATASET IS SPLIT BY SUBJECTS, WHERE THE ACTIONS OF SOME SUBJECTS ARE USED FOR TRAINING AND THOSE OF THE REMAINING SUBJECTS ARE USED FOR TESTING. THE RESULTS OF "OURS" ARE OBTAINED BY RUNNING OUR METHOD FIVE TIMES. BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED

	Datasets	UWA3D VIEW3 [36]				NW-UCLA [37]				NTU RGB+D 60 CS [38]			
	%Labels	5%	10%	20%	50%	5%	15%	30%	40%	1%	2%	5%	10%
	#Labels	25	50	100	250	50	150	300	400	400	800	2000	4000
Baselines	Discriminator-based Selection [59]	19.3	28.7	40.2	53.8	47.7	71.8	76.6	80.5	34.9	39.5	53.8	60.4
Baselines	Core Set [58]	21.5	29.9	40.3	52.6	57.3	69.4	77.3	80.6	17.6	23.1	37.0	49.6
Baselines	Consistency-based AL under Augmentation [60]	21.9	30.2	40.8	56.2	48.2	65.8	77.0	81.7	20.7	33.5	50.4	60.2
Baselines	Uniform Sampling [17]	23.4	30.5	42.1	53.4	52.3	70.4	78.4	80.8	34.6	43.8	56.3	61.2
Baselines	AL-SAR [17]	25.7	35.3	45.7	53.9	61.0	75.9	82.5	84.1	38.8	47.6	57.9	63.7
	Ours	28.9 ± 0.2	39.1 ± 0.1	49.1 ± 0.3	58.7 ± 0.1	64.3 ± 0.2	79.1 ± 0.1	86.0 ± 0.2	87.9 ± 0.2	41.1 ± 0.3	51.3 ± 0.1	60.2 ± 0.3	66.9 ± 0.2

augmentation selects samples based on the consistency of augmented skeleton sequences.

Besides, following [17], we compare our method with three representative semi-supervised skeleton-based 3D action recognition methods: ASSL [27], MS^{2}L [28], and SC3D [29]. It is important to note that these methods assume access to a pre-defined labeled set and do not address the problem of active sample selection.

C. Implementation Details

For fair comparison, we follow [17] and adopt an encoder-decoder architecture to encode skeleton data. Specifically, our encoder consists of three-layer bi-GRU cells with 1024 hidden units in each direction. The hidden states from both directions are concatenated into a 2048-dimensional latent representation, which is then fed into the decoder. The decoder is a uni-directional GRU with a hidden size of 2048. For a fair comparison, we follow [17] and use the Adam optimizer [56] to train the action recognizer. Both the encoder–decoder, optimized with a reconstruction loss, and the action recognizer, optimized with a classification loss, are trained simultaneously. The learning rate is initialized to 10^{-4} and decayed by a factor of 0.95 every 10 epochs on the UWA3D and NW-UCLA datasets, and every 3 epochs on the NTU RGB+D 60 dataset. Our informative sample selection model (ISSM) is a lightweight three-layer MLP. For a fair comparison, AL-SAR and several active learning baselines adopt the same three-layer MLP as the action recognizer, consistent with our method. We also use the Adam optimizer [56] as its optimizer. The learning rate for the ISSM is set to 1×10^{-4} across all three datasets: UWA3D, NW-UCLA, and NTU RGB+D 60. Specifically, our model is trained for 100 epochs on the UWA3D and NW-UCLA datasets, and for 300 epochs on the NTU RGB+D dataset. Our ISSM is trained for 50 epochs on the UWA3D and NW-UCLA datasets, and for 100 epochs on the NTU RGB+D 60 dataset. The model is pre-trained in advance and kept frozen during our Semi-supervised 3D Action Recognition via Active Learning task.

D. Main Experiments

In the previous subsection, we introduce the evaluation datasets and metric, the baselines for comparison, and the implementation details of our method. In this subsection, we present the comparison results of our method against state-of-the-art (SOTA) active learning baselines on the three datasets, i.e., UWA3D, NW-UCLA, and NTU RGB+D 60.

TABLE II

WE COMPARE THE PERFORMANCE OF ACTION RECOGNIZERS TRAINED ON SAMPLES OBTAINED THROUGH RANDOM SELECTION, AS USED IN PREVIOUS SEMI-SUPERVISED ACTION RECOGNITION METHODS (MS^{2}L [28], AND SC3D [29]), WITH THOSE TRAINED ON SAMPLES SELECTED BY OUR METHOD AND AL-SAR [17],ON THE WIDELY-USED 3D ACTION RECOGNITION DATASETS NTU RGB+D 60 [14] UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES. "%LABELS" DENOTES THE PROPORTION OF LABELED SAMPLES RELATIVE TO THE TOTAL NUMBER OF TRAINING SAMPLES IN THE DATASET, WHILE "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. BEST RESULTS ARE HIGHLIGHTED

Datasets	NTU RGB+D 60 CS [38]			
%Labels	1%	2%	5%	10%
#Labels	400	800	2000	4000
MS^{2}L [29]	33.1	40.3	57.8	65.2
MS^{2}L [29] + AL-SAR [17]	35.3	42.8	59.7	67.0
MS^{2}L [29] + Ours	38.0	45.9	61.2	69.8
SC3D [30]	35.7	41.2	59.6	65.9
SC3D [30] + AL-SAR [17]	36.6	42.8	61.0	67.3
SC3D [30] + Ours	39.2	45.6	63.8	69.5

1) Comparison With SOTA Active Learning Methods: As shown in Table I, we compare our method with previous state-of-the-art active learning approaches on commonly used 3D action recognition datasets under different numbers of labeled samples. As can be seen in Table I, our method significantly outperforms previous SOTA active learning methods. For example, on the UWA3D dataset with 50 labeled samples, our method achieves an accuracy of 39.1%, surpassing the previous best method AL-SAR by 3.8%.

TABLE III

ABLATION STUDY ON OUR DESIGNED STATE (DISTRIBUTION GAP) AND ACTION REPRESENTATIONS (MARGINAL INDEX AND SKELETON SEQUENCE REPRESENTATIVENESS) ACROSS THREE COMMONLY USED 3D ACTION RECOGNITION DATASETS (UWA3D [35], NW-UCLA [36], AND NTU RGB+D 60 [14]) UNDER VARYING PROPORTIONS OF LABELED SAMPLES. "%LABELS" DENOTES THE PROPORTION OF LABELED SAMPLES RELATIVE TO THE TOTAL NUMBER OF TRAINING SAMPLES IN THE DATASET, WHILE "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. BEST RESULTS ARE HIGHLIGHTED IN THE TABLE

Type	Datasets	UWA3D VIEW3 [36]		NW-UCLA [37]		NTU RGB+D 60 CS [38]	
	%Labels	5%	20%	5%	30%	1%	5%
	#Labels	25	100	50	300	400	2000
State	Ours w/o Distribution Gap	18.1	38.7	45.6	73.1	15.9	38.1
	Ours w/o Budget Consumption Ratio B^C	25.1	45.5	52.8	79.0	36.3	54.4
Action	Ours w/o Marginal Index	24.1	45.3	52.1	78.7	35.6	54.2
	Ours w/o Skeleton Sequence Representativeness	25.6	46.0	53.9	79.1	37.0	54.9
	Ours	28.9	49.1	64.3	86.0	41.1	60.2

TABLE IV

ABLATION STUDY ON OUR DESIGNED HYPERBOLIC REPRESENTATION ON THREE COMMON 3D ACTION RECOGNITION DATASETS (UWA3D [35], NW-UCLA [36], AND NTU RGB+D 60 [14]) UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES. "%LABELS" DENOTES THE PROPORTION OF LABELED SAMPLES RELATIVE TO THE TOTAL NUMBER OF TRAINING SAMPLES IN THE DATASET, WHILE "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. BEST RESULTS ARE HIGHLIGHTED

Datasets	UWA3D VIEW3 [36]		NW-UCLA [37]		NTU RGB+D 60 CS [38]	
%Labels	5%	20%	5%	30%	1%	5%
#Labels	25	100	50	300	400	2000
Ours w/o Hyperbolic Representation	26.8	47.1	60.5	82.3	39.0	57.8
Ours	28.9	49.1	64.3	86.0	41.1	60.2

TABLE V

ABLATION STUDY OF OUR DESIGNED META TUNING ON THREE COMMON 3D ACTION RECOGNITION DATASETS (UWA3D [35], NW-UCLA [36], AND NTU RGB+D 60 [14]) UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES. "TIME" DENOTES THE DURATION REQUIRED FOR THE MODEL TO CONVERGE FROM THE START OF TRAINING. "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. ✓ AND ✗ REPRESENT WHETHER META-TUNING IS PERFORMED OR NOT, RESPECTIVELY

Datasets	#Labels	Meta Tuning	Accuracy (%)	Time (hour)
UWA3D VIEW3 [36]	25	✓	28.8	0.4
	25	✗	28.9	0.7
	100	✓	48.7	1.0
	100	✗	49.1	2.5
NW-UCLA [37]	50	✓	64.2	0.4
	50	✗	64.3	0.8
	300	✓	85.8	1.7
	300	✗	86.0	3.3
NTU RGB+D 60 CS [38]	400	✓	40.9	0.6
	400	✗	41.1	1.0
	2000	✓	59.9	2.5
	2000	✗	60.2	5.1

Our method also demonstrates strong robustness across datasets of different scales, including the relatively smaller UWA3D and NW-UCLA datasets, as well as the larger NTU RGB+D 60 dataset. For example, on the large-scale NTU RGB+D 60 dataset with 800 labeled samples, our method achieves an accuracy of 51.3%, outperforming the previous SOTA method AL-SAR, which achieves 47.6%, by 3.7%. In addition, on the moderately sized NW-UCLA dataset, our method consistently outperforms the prior SOTA method. For instance, under the setting of 400 labeled samples, our method achieves 87.9% accuracy, while AL-SAR only reaches 84.1%, resulting in a 3.8% improvement. Moreover, as shown in Table I, our method also exhibits strong robustness under varying numbers of labeled samples within the same dataset. For example, on the NW-UCLA dataset, our method achieves 64.3% accuracy with only 50 labeled samples, and maintains high performance with more labeled samples, reaching 87.9% accuracy with 400 samples.

In addition, we combine our method and AL-SAR with two state-of-the-art semi-supervised skeleton-based action

recognition methods, MS^{2}L and SC3D, to select training samples, and evaluate their performance on the NTU RGB+D 60 dataset (see Table II). Experimental results show that MS^{2}L and SC3D achieve better action recognition performance when trained on samples selected by our method. For example, under the 4,000 labeled samples setting on the NTU RGB+D 60 dataset, MS^{2}L + AL-SAR achieves an accuracy of 67.0%, while MS^{2}L + Ours achieves 69.8%.

Overall, our method demonstrates strong robustness across different datasets and labeling ratios, consistently outperforming previous SOTA methods. We attribute this to the use of a theoretically grounded MDP framework for training our informative sample selection model, enabling more intelligent sample selection.

E. Ablation Studies

In the previous subsection, we present the main experiments of our work, comparing our method with previous SOTA active learning methods on three widely used skeleton-based action recognition datasets.

In this subsection, we first follow AL-SAR [17] to verify the performance gains brought by the active learning framework in our method. In addition, we analyze the impact of our state and action space designs, the projection of representations from Euclidean space to hyperbolic space, and the effectiveness of the meta tuning strategy. Finally, we analysis the generalization of our method.

We follow AL-SAR [17] to conduct our ablation studies under the settings of 25 and 100 labeled samples on the UWA3D dataset, 50 and 300 labeled samples on the NW-UCLA dataset, and 400 and 2000 labeled samples on the NTU RGB+D 60 dataset.

1) Impact of State and Action Space: As shown in Table III, we conduct ablation studies on the distribution gap in the state and the marginal index and skeleton sequence representativeness in the action across three commonly used skeleton-based action recognition datasets. In the setting "Ours w/o Distribution Gap", the state is randomly generated. The results show that removing the Distribution Gap from the state leads to a 10.8% drop in accuracy, demonstrating its effectiveness in helping our method identify informative skeleton sequences and train a better action recognizer.

From the experimental results of "Ours w/o Budget Consumption Ratio B^C", we observe that incorporating B^C into the state leads to further performance improvements. We attribute this to the ability of B^C to help the method better recognize the current learning stage of the model, which in turn facilitates the selection of samples that are more suitable for the corresponding training stage.

Furthermore, the ablation results on the action space confirm the effectiveness of both the Marginal Index and Skeleton Sequence Representativeness. These two factors are shown to be complementary and jointly contribute to the performance improvement of our method.

2) Impact of Hyperbolic Representation: To better classify human actions based on skeletal data, we design a Hyperbolic Representation tailored to the tree-like structure of human skeletons for both state and action representations. Here, we conduct an ablation study on this Hyperbolic Representation. As shown in Table IV, applying the Hyperbolic Representation leads to improved performance, demonstrating the effectiveness of this representation space.

TABLE VI

QUANTITATIVE COMPARISONS WITH REPRESENTATIVE ACTIVE LEARNING METHODS (UNIFORM SAMPLING [17], CORE SET SELECTION [53], DISCRIMINATOR-BASED SELECTION [54], AND CONSISTENCY-BASED ACTIVE LEARNING UNDER AUGMENTATION [55], AND AL-SAR [17]) ON THREE COMMON 3D ACTION RECOGNITION DATASETS (UWA3D [35], NW-UCLA [36], AND NTU RGB+D 60 [14]) UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES. "%LABELS" DENOTES THE PROPORTION OF LABELED SAMPLES RELATIVE TO THE TOTAL NUMBER OF TRAINING SAMPLES IN THE DATASET, WHILE "#LABELS" INDICATES THE ABSOLUTE NUMBER OF LABELED SAMPLES. "VIEW3" REFERS TO USING THE THIRD CAMERA VIEW OF THE MULTI-VIEW UWA3D DATASET AS THE TESTING SET, WHILE THE REMAINING VIEWS ARE USED FOR TRAINING. "CS" STANDS FOR "CROSS-SUBJECT", MEANING THAT THE DATASET IS SPLIT BY SUBJECTS, WHERE THE ACTIONS OF SOME SUBJECTS ARE USED FOR TRAINING AND THOSE OF THE REMAINING SUBJECTS ARE USED FOR TESTING. THE RESULTS OF "OURS" ARE OBTAINED BY RUNNING OUR METHOD FIVE TIMES

	Datasets	NW-UCLA [37]			
	%Labels	5%	15%	30%	40%
	#Labels	50	150	300	400
Baselines	Discriminator-based Selection [59]	47.7	71.8	76.6	80.5
	Core Set [58]	57.3	69.4	77.3	80.6
	Consistency-based AL under Augmentation [60]	48.2	65.8	77.0	81.7
	Uniform Sampling [17]	52.3	70.4	78.4	80.8
	AL-SAR [17]	61.0	75.9	82.5	84.1
	Ours w/o retraining	63.2	77.8	85.3	86.5
	Ours w/ retraining	64.3	79.1	86.0	87.9

3) Impact of Meta Tuning: To accelerate the deployment of our method in real-world scenarios, we introduce a meta tuning strategy based on meta-learning [34]. Here, we conduct an ablation study to evaluate its effectiveness (see Table V). In the table, "Time" denotes the convergence time required for training. As shown by the results, our meta tuning significantly speeds up model convergence. For example, under the NTU RGB+D 60 dataset with 2000 labeled samples, "✓" converges in only 2.5 hours, compared to 5.1 hours for "✗", achieving nearly a 50% reduction in training time with comparable accuracy. Moreover, across all three datasets, "✓" consistently achieves similar accuracy to "✗" while reducing convergence time by roughly half, demonstrating the robustness of the proposed meta tuning strategy.

4) Generalization Ability: To demonstrate the generalization capability of our method, we add new experiments. Specifically, the newly added experiments evaluate the generalization of our Informative Sample Selection Model by training it with 100 samples on the UWA3D VIEW3 dataset and testing it on the NW-UCLA dataset. As shown in the experimental results (see VI), our method surpasses previous approaches even without being trained on the NW-UCLA dataset, demonstrating its strong generalization ability.

V. CONCLUSION

In this paper, we propose a novel perspective for the semi-supervised 3D action recognition via active learning task by

formulating it as a Markov Decision Process (MDP), aiming to address the limitation of previous margin-based selection strategy that may fail to select samples that effectively enhance the performance of the action recognizer. Specifically, we leverage the theoretically grounded MDP framework to train an informative sample selection model that intelligently selects representative samples for annotation. To enhance the representational capacity of the state-action pairs in our MDP framework, we map them from the Euclidean space to hyperbolic space. Moreover, we explore a meta tuning strategy based on meta-learning to accelerate the deployment of our method in real-world scenarios. Extensive experiments are conducted on three widely used 3D action recognition datasets: UWA3D, North-Western UCLA, and NTU RGB+D 60. The results show that our method yields greater improvements in action recognizer performance compared to prior approaches and exhibits strong generalization capability.

ACKNOWLEDGMENT

The numerical calculation was supported by Super-Computing System in Super-Computing Center of Wuhan University.

REFERENCES

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

[2] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 55–63.

[3] Y. Zhu, H. Shuai, G. Liu, and Q. Liu, "Multilevel spatial–temporal excited graph network for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 32, pp. 496–508, 2023.

[4] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 329–345.

[5] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.

[6] I. Rodomagoulakis et al., "Multimodal human action recognition in assistive human-robot interaction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2702–2706.

[7] R. Yue, Z. Tian, and S. Du, "Action recognition based on RGB and skeleton data sets: A survey," *Neurocomputing*, vol. 512, pp. 287–306, Nov. 2022.

[8] W. Lin, X. Ding, Y. Huang, and H. Zeng, "Self-supervised video-based action recognition with disturbances," *IEEE Trans. Image Process.*, vol. 32, pp. 2493–2507, 2023.

[9] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal VLAD for video action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2799–2812, Jun. 2019.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 1, pp. 51–61, Feb. 2015.

[11] Z. Xu et al., "Semisupervised discriminant multimanifold analysis for action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2951–2962, Oct. 2019.

[12] S. Guan, X. Yu, W. Huang, G. Fang, and H. Lu, "DMMG: Dual min-max games for self-supervised skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 395–407, 2024.

[13] W. Myung, N. Su, J.-H. Xue, and G. Wang, "DeGCN: Deformable graph convolutional networks for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 2477–2490, 2024.

[14] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1010–1019.

[15] H.-B. Zhang et al., "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[16] L. Xu, C. Lan, W. Zeng, and C. Lu, "Skeleton-based mutually assisted interacted object localization and human action recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 4415–4425, 2023.

[17] J. Li, T. Le, and E. Shlizerman, "AL-SAR: Active learning for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16966–16974, Nov. 2024.

[18] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[19] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[20] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[21] Z. Zhang, C. Zhou, and Z. Tu, "Distilling inter-class distance for semantic segmentation," 2022, *arXiv:2205.03650*.

[22] J. Zhang, H. P. H. Shum, J. Han, and L. Shao, "Action recognition from arbitrary views using transferable dictionary learning," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4709–4723, Oct. 2018.

[23] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5479–5490, Nov. 2016.

[24] Z. Zhang, L. G. Foo, H. Rahmani, J. Liu, and D. W. Soh, "Performing defocus deblurring by modeling its formation process," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2025, pp. 5791–5801.

[25] Z. Tu, Y. Liu, Y. Zhang, Q. Mu, and J. Yuan, "DTCM: Joint optimization of dark enhancement and action recognition in videos," *IEEE Trans. Image Process.*, vol. 32, pp. 3507–3520, 2023.

[26] Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning for video-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 4104–4116, 2022.

[27] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3D action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 35–51.

[28] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2490–2498.

[29] F. M. Thoker, H. Doughty, and C. G. M. Snoek, "Skeleton-contrastive 3D action representation learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1655–1663.

[30] J. Gong, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Meta agent teaming active learning for pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11069–11079.

[31] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 2094–2100.

[32] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. B. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, 2016, pp. 3682–3690.

[33] H. Qu, Y. Cai, and J. Liu, "LLMs are good action recognizers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 33, Jun. 2024, pp. 18395–18406.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.

[35] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 742–757.

[36] J. Wang, X. Nie, X. Yin, Y. Wu, and S. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2649–2656.

[37] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with deeply transferred motion vector CNNs," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2326–2339, May 2018.

[38] Z. Zhang, L. Xu, D. Peng, H. Rahmani, and J. Liu, "Diff-tracker: Text-to-image diffusion models are unsupervised trackers," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 319–337.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7444–7452.

[40] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.
[41] L. Wang and P. Koniusz, "3Mformer: Multi-order multi-mode transformer for skeletal action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5620–5631.
[42] Z. Zhang et al., "Visual prompting for one-shot controllable video editing without inversion," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Jun. 2025, pp. 7784–7794.
[43] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13359–13368.
[44] L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, "Unified pose sequence modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 13019–13030.
[45] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang, "Semi-supervised image-to-video adaptation for video action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 960–973, Apr. 2017.
[46] A. Singh et al., "Semi-supervised action recognition with temporal contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10384–10394.
[47] 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 *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1.
[48] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3D action recognition," 2020, *arXiv:2007.05934*.
[49] M. Lauri, D. Hsu, and J. Pajarinen, "Partially observable Markov decision processes in robotics: A survey," *IEEE Trans. Robot.*, vol. 39, no. 1, pp. 21–40, Feb. 2023.
[50] O.-E. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 5350–5360.
[51] H. Qu, X. Hui, Y. Cai, and J. Liu, "LMC: Large model collaboration with cross-assessment for training-free open-set object recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 46491–46504.
[52] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Proc. Eur. Conf. Mach. Learn.* Cham, Switzerland: Springer, 2006, pp. 413–424.
[53] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.
[54] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5971–5980.
[55] M. Gao, Z. Zhang, G. Yu, S. Ö. Arik, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 510–526.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

Zhigang Tu (Senior Member, IEEE) received the Ph.D. degree from Wuhan University, China, in 2013, and the Ph.D. degree from Utrecht University, The Netherlands, in 2015. From 2015 to 2016, he was a Postdoctoral Researcher with Arizona State University, USA. From 2016 to 2018, he was a Research Fellow with Nanyang Technological University, Singapore. He is currently a Professor with Wuhan University. He has co-/authored more than 80 papers in international SCI-indexed journals and conferences. His current research interests include computer vision, image processing, video analytics, machine learning, motion estimation, human action and gesture recognition, and anomaly event detection. He received the Best Student Paper Award at the Fourth Asian Conference on Artificial Intelligence Technology, and one of the three best reviewers awards for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2022. He is the first organizer of ACCV2020 Workshop on MMHAU, Japan. He was the Area Chair of AAAI2023/2024 and VCIP2022, an Associate Editor of the SCI-indexed journal *The Visual Computer* (IF is 3.5), and a Guest Editor of *Journal of Visual Communications and Image Representation* (IF is 2.6).

Zhengbo Zhang (Member, IEEE) received the bachelor's and master's degrees in engineering from Wuhan University. He is currently pursuing the Ph.D. degree in information systems technology and design with Singapore University of Technology and Design. His current research interests are in computer vision and machine learning.

Jia Gong received the B.Eng. degree in opto-electronic engineering from Chongqing University, China, and the Ph.D. degree in information systems technology and design from Singapore University of Technology and Design. He is a Researcher with Shanghai Academy of Artificial Intelligence for Science. His research interests span human pose estimation, human mesh reconstruction, digital avatars, and active learning.

Junsong Yuan (Fellow, IEEE) received the B.Eng. degree from the Huazhong University of Science Technology in 2002, the M.Eng. degree from the National University of Singapore in 2005, and the Ph.D. degree from Northwestern University in 2009. He is a Professor and the Director of the Visual Computing Laboratory, Department of Computer Science and Engineering, State University of New York, Buffalo (UB), USA. Before that, he was an Associate Professor (2015–2018) and a Nanyang Assistant Professor (2009–2015) with Nanyang Technological University (NTU), Singapore. His research interests include computer vision, pattern recognition, video analytics, and large-scale visual search and mining. He is a fellow of IAPR. He received the Best Paper Award from IEEE TRANSACTIONS ON MULTIMEDIA, the Nanyang Assistant Professorship from NTU, and the Outstanding EECS Ph.D. Thesis Award from Northwestern University. He was the Program Co-Chair of IEEE Conference on Multimedia Expo (ICME'18/2022/2024); and the Area Chair of CVPR, ICCV, ECCV, and ACM MM. He was an Elected Senator at NTU and UB. He served as an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESS, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY; and a Senior Area Editor for *Journal of Visual Communications and Image Representation*.

Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010. He is a Professor with the School of Computer Science, Wuhan University. He has over 80 research articles published in the journals of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *ISPRS Journal of Photogrammetry and Remote Sensing*. More than 30 of them are ESI hot articles or highly cited articles. His major research interests include pattern recognition, hyperspectral image processing, machine learning, and signal processing.