Local Dimension Enhancement Representation Learning for Skeleton-Based Action Segmentation

Shaofan Sun, Lilang Lin, *Graduate Student Member, IEEE*, Jiahang Zhang, Ling-Yu Duan, *Member, IEEE*, and Jiaying Liu, *Fellow, IEEE*

***Abstract*—Most existing self-supervised learning methods for skeleton-based temporal action segmentation (TAS) fail to capture the short-term motion semantics essential for dense frame-level prediction, as they typically learn representations that are either too coarse or motion-insensitive. This issue is reflected in local dimension collapse, which highlights the limitations of current approaches and suggests directions for improvement. Specifically, to address the issue of local dimension collapse for self-supervised learning in TAS, we propose the Local Dimension Enhancement (LoDE) framework, which introduces the local effective rank (LER) as a metric to measure and a learning objective to reduce this collapse. A new fine-grained representation scale, termed a motion unit, is defined as a temporal clip of consecutive skeleton frames to model skeleton data. Centered on this representation scale, we analyze existing methods (sequence-scale and frame-scale learning) with the tool of LER and theoretically demonstrate that introducing motion unit-scale learning is essential to alleviate local dimension collapse. Inspired by our theoretical insights, we design a multi-scale semantics module that integrates frame-, sequence-, and motion unit-scale learning, with LER-based regularization to enrich local representation diversity. These designs effectively alleviate local dimension collapse and lead to significant improvements in TAS, as evidenced by LoDE's superior performance over state-of-the-art methods on three large-scale untrimmed datasets: PKUMMD, TSU, and BABEL. Our project website is available at https://carefreesun.github.io/LoDE_TIP_2026/**

***Index Terms*—Skeleton-based action segmentation, representation learning, dimension collapse.**

I. Introduction

HUMAN action understanding has witnessed remarkable advances in recent years through the use of 3D skeleton data, which offers lightweight, compact, and privacy-preserving representations [1], [2], [3], [4], [5], [6], [7], [8], [9]. However, the success of prevailing methods heavily depends on fully supervised training, which requires costly and labor-intensive annotations. Moreover, due to the intrinsic characteristics of human actions such as diversity, contextual complexity, and long-tailed distribution, these methods often fall short in effectively modeling complex real-world scenarios. These critical limitations necessitate a shift towards more scalable paradigms, making self-supervised learning (SSL) an increasingly vital approach for learning generalizable representations from large amounts of unlabeled skeleton data [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24].

SSL for skeleton has been well studied in action recognition [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], which aims to learn meaningful representations from unlabeled skeleton data by solving pretext tasks and then use the pre-trained model to predict a single action class for each trimmed sequence. However, limited attention has been paid to a more challenging action understanding task named temporal action segmentation (TAS) [21], [22], [23], [24], which directly works with untrimmed videos and is more practical for real-world applications. In addition to recognizing the action categories, skeleton-based TAS requires localizing the temporal boundaries for untrimmed sequences composed of multiple actions and transitions. This presents significant challenges for SSL methods in modeling local representations, *i.e.*, learning highly distinguishable representations for adjacent frames or short-term motions, to perceive subtle changes.

Specifically, existing methods can be generally divided into sequence-scale and frame-scale learning. Sequence-scale learning methods, *e.g.*, contrastive learning [10], [11], [12], [18], encode each skeleton sequence into a global representation and perform inter-sequence constraints, which lack the modeling of local relationships and discrepancies. This restricts the models' sensitivity to fine-grained motion changes. Frame-scale learning methods [14], [15], [17], [25] learn representations via reconstructing original or transformed skeleton, and thus focus on the detailed information within each sequence and perceive little inter-class semantics. This leads to less meaningful local representations and degrades the boundary prediction performance.

More formally, the seemingly different limitations of sequence-scale and frame-scale learning originate from a deeper, shared fundamental problem called dimension collapse [26], [27], [28]. In SSL, dimension collapse is one of the intrinsic issues, where learned representations lie in a low-dimensional subspace, failing to capture the full data manifold and degrading downstream perception performance. Specifically, since TAS requires diverse local representations within sequences, we focus on the issue of **local dimension**

Received 8 October 2025; revised 24 February 2026; accepted 29 March 2026. Date of publication 14 April 2026; date of current version 17 April 2026. This work was supported in part by Beijing Major Science and Technology Project under Contract Z251100008425023. The associate editor coordinating the review of this article and approving it for publication was Prof. Leida Li. *(Corresponding author: Jiaying Liu.)*

The authors are with Peking University, Beijing 100871, China (e-mail: carefree_sun@stu.pku.edu.cn; linlilang@pku.edu.cn; zjh2020@pku.edu.cn; lingyu@pku.edu.cn; liujiaying@pku.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2026.3682105, provided by the authors.

Digital Object Identifier 10.1109/TIP.2026.3682105

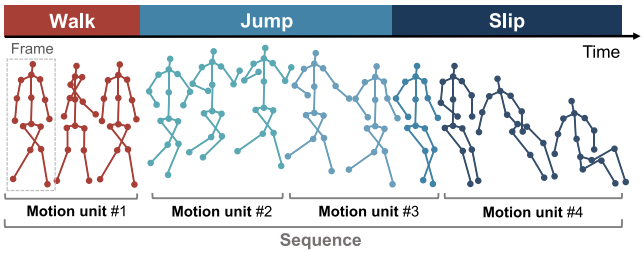


Fig. 1. Visualization of motion unit compared with frame and sequence.

collapse, as formalized by LDReg [29], where representations collapse within localized regions. This perspective enables a quantitative analysis of existing SSL frameworks' limitations in modeling local representations and guides corresponding improvements.

Specifically, to mitigate local dimension collapse for skeleton-based TAS, we propose a Local Dimension Enhancement (LoDE) learning framework. First, to quantify local dimension collapse, we theoretically derive **local effective rank** (LER) from the perspective of manifold dimension [30] as a measure for the dimensionality of the latent space within a local region. Moreover, beyond typical modeling scales of sequence and frame, we introduce an intermediate scale named **motion unit**, *i.e.*, a clip covering consecutive frames of certain joints, to model local representations for skeleton data. We present a visual illustration of different scales in Figure 1. If we compare a sequence to a sentence, frames are like the letters, and then motion units represent the words, which are finer than sequences while containing more motion semantics than frames. Therefore, motion units are more suitable for skeleton representation learning and analysis.

With the proposed LER, we analyze the SSL methods with learning objectives of different temporal scales, *i.e.*, sequence, frame, and motion unit scale. We demonstrate that both frame-scale and sequence-scale learning suffer severely from local dimension collapse, while motion unit-scale learning can help enhance LER by directly improving local representation diversity. Based on the analysis, we propose a multi-scale action semantics learning module centered on the motion unit scale. In addition to motion unit-scale learning, we integrate finer-grained frame-scale learning to capture action details and coarser-grained sequence-scale learning to perceive long-range semantics. Besides, we integrate LER to regularize the latent manifold. By uniformizing the singular value distribution of the local representation matrix, it further diversifies the representations, which enhances the model's ability to perceive subtle motion changes for precise boundary localization.

Our learning framework effectively alleviates local dimension collapse and improves the downstream TAS performance. Extensive experiments on PKUMMD [31], TSU [32], and BABEL [33] datasets demonstrate the remarkable performance of our model compared to the state-of-the-art methods.

Our contributions can be summarized as follows:

- We propose a novel local dimension enhancement learning framework to address the local dimension collapse problem in skeleton-based TAS. We quantify the problem with a new proposed manifold dimension measure LER. During training, the framework incorporates motion unit-scale learning and LER regularization to generate discriminative local representations.
- We theoretically analyze local dimension collapse using LER, showing that lower LER correlates with more severe collapse, greater information loss, and poorer TAS performance. These insights motivate incorporating LER as a regularization to learn information-preserving representations.
- We present that motion unit-scale learning improves local dimension indicated by LER. Based on it, we integrate sequence and frame scales to construct multi-scale action semantics learning for more diverse local representations. Finer frame-scale learning refines motion units with details, and coarser sequence-scale learning enhances global semantic perception.

The remainder of the paper is organized as follows. In Sec. II, prior related works are introduced. Sec. III presents our proposed LER measure along with its theoretical analysis. Sec. IV elaborates the architectural details of the proposed Local Dimension Enhancement learning framework. In Sec. V, we present the experimental results and the corresponding analysis. Sec. VI concludes our work.

II. Related Works

In this section, we briefly introduce the preliminaries of temporal action segmentation, skeleton-based action representation learning, and alleviation of dimension collapse, and review the related prior works.

A. Temporal Action Segmentation

Temporal action segmentation (TAS) focuses on frame-scale action classification for untrimmed videos. Commonly, pre-trained feature extractors are applied to the videos to obtain dense representations, which are then post-processed by segmentation models into frame-scale labels or temporal segments. Many designs focus on the architecture of the segmentation model. JCRRNN [34] uses recurrent neural networks to perform regression. MS-TCT [35] leverages multi-scale knowledge fusion to capture long-term semantics. DiffAct [36] employs extracted representations as conditional inputs, utilizing a diffusion model to denoise labels for segmentation. To further improve accuracy and generalizability, more and more works incorporate pre-training strategies for feature extractors to generate high-quality representations. GRU-GD [37] leverages multi-modal data for abundant information. LAC [21] enhances the diversity of motion patterns through synthetic action generation. BID [23] explicitly incorporates boundary prediction during pre-training to improve localization accuracy. SCS [24] performs contrastive learning on concatenated trimmed sequences for model training. Building on these insights, our work focuses on the pre-training paradigm, aiming to produce highly informative and discriminative representations without relying on external data.

B. Sequence-Scale Skeleton Representation Learning

Skeleton-based representation learning methods aim to learn informative representations for skeleton sequences with unlabeled data and benefit downstream tasks. Among these, sequence-scale learning has emerged as a prominent approach. It operates by encoding an entire skeleton sequence into a single global representation. The learning objective is then defined by inter-sequence relations, compelling the model to capture coarse-grained action semantics suitable for tasks like action recognition. The most popular sequence-scale learning paradigm is contrastive learning, which operates by maximizing the similarity between positive pairs while minimizing it for negative pairs, thereby structuring a well-separated representation space. Rao et al. [38] applied a MoCo-based [39] framework and designed a series of data augmentations tailored for skeleton data. Guo et al. [10] applied stronger augmentations to explore more action patterns and benefit contrastive learning. Zhang et al. [11] proposed a hierarchical learning framework to maintain the consistency of representation semantics when applying strong augmentations. Lin et al. [16] proposed to integrate equivariant and invariant data transformations to encourage capturing augmentation-related information. Although these methods can learn the sequence-scale semantics, they lack modeling of short-term motion patterns, which limits their applicability to TAS tasks.

C. Frame-Scale Skeleton Representation Learning

To mine short-term motion patterns and capture fine-grained action semantics in skeleton sequences more effectively, frame-scale learning methods are proposed. LAC [21] applies frame-scale contrastive learning to distinguish small changes in the action. Based on masked autoencoders (MAEs) [40], SkeletonMAE [14] learns representations by reconstructing the full skeleton from masked joints. MAMP [15] further discovers that reconstructing motion can significantly improve the representation quality and downstream performance. MacDiff [17] uses a diffusion model for reconstruction and obtains more robust and generalizable action representations. While these methods excel at capturing local details, their reliance on reconstruction from low-dimensional inputs overlooks the inter-sequence semantics and inherently limits the dimension of the learned representations, failing to discover a latent space rich enough for complex semantic distinctions. To leverage the advantages of sequence-scale learning and frame-scale learning to perceive richer semantics, hybrid learning frameworks are proposed. MS^{2}L [41] directly combines multiple tasks of sequence-scale contrastive learning, frame-scale motion prediction and Jigsaw puzzle to learn multi-scale dynamics. PCM^3 [13] integrates contrastive learning and masked skeleton modeling paradigms in a mutually beneficial manner to better fuse semantics from different granularities. However, simply combining sequence-scale and frame-scale objectives often neglects the intermediate temporal structures, leading to a semantic gap where the compositional nature of complex actions is not explicitly modeled. To overcome the limitations, our method additionally proposes a motion unit-scale learning paradigm to integrate multi-scale learning.

D. Alleviation of Dimension Collapse

Dimension collapse is a prevalent issue in SSL. The worst case is *complete collapse*, where all the samples are mapped to a constant representation, resulting in no meaningful information. To address this problem, many contrastive methods [39], [42], non-contrastive methods [43], [44] and autoencoder-like methods [40] are proposed. However, these methods still suffer from a more subtle form of degeneracy known as *dimension collapse*, where the subspace spanned by the representations is low-dimensional. Recent works have made great endeavors in analyzing and alleviating dimension collapse. Some works analyze intrinsic dimension based on the spectrum of the representation space, addressing dimension collapse through approaches such as regularization loss [27], model architecture [45], and data sampling [28], respectively. Fang et al. [26] applied a uniformity metric based on Wasserstein distance to constrain model training. However, most of the methods focus on the global structure of the representations. From a more fine-grained perspective, LDReg [29] tackles the *local dimension collapse* problem, where representations can collapse locally while spanning a high-dimensional global latent space. This inspires us to examine the local dimension collapse at the motion unit scale for skeleton-based TAS.

III. Analysis on Local Dimension Collapse

In this section, we first introduce the LER measure for evaluating the local intrinsic dimension, and theoretically derive its negative relation with the extent of local dimension collapse. With the tool of LER, we further analyze different skeleton-based representation learning methods in terms of temporal modeling scales. Then we explore an effective method integrating a novel modeling scale, *i.e.*, motion unit, to mitigate local dimension collapse for skeleton-based TAS.

A. Local Effective Rank Measure

In standard terminology, the dimensionality of a data manifold or latent space is referred to as its *intrinsic dimension*. Local dimension collapse occurs when the intrinsic dimension becomes significantly reduced within localized regions, which enables its quantification through intrinsic dimension measurement. Generally, the measures to evaluate intrinsic dimension can be divided into *fractal dimension* and *manifold dimension* [30]. The fractal dimension is defined by the following formulation [46]:

Definition 1: Consider a real-valued function F that remains non-zero over some open interval containing $r \in \mathbb{R}, r \neq 0$. The intrinsic dimension of F at r is defined as:

$$\mathrm{ID}_F(r) \triangleq \lim_{\epsilon \to 0} \frac{\log\{F[(1+\epsilon)r]/F(r)\}}{\log[(1+\epsilon)r/r]}. \tag{1}$$

However, the latent spaces to analyze are typically discrete rather than continuous. This poses challenges for fractal dimension estimation since Eq. (1) fundamentally relies on a limit-based operation, requiring either continuity or dense discrete sampling to approximate the limit. Unfortunately, such dense sampling is often infeasible within a localized region, where the available representations are limited in

quantity. Consequently, fractal dimension measures tend to exhibit instability when measuring local intrinsic dimension.

To develop a more stable local intrinsic dimension measure, we propose the *local effective rank* derived from the *manifold dimension*. The definitions of the embedding manifold and the corresponding manifold dimension [30] are given as follows:

Definition 2: Let $d < D$ and let Ω be a compact open set in $\mathbb{R}^d$. Assume that span $\left\{\Omega - \int_\Omega d\mu\right\} = \mathbb{R}^d$ and $\phi : \Omega \to \mathbb{R}^D$ is a smooth function. The set $\mathcal{X} = \phi(\Omega)$ is called an embedding manifold with manifold dimension of d.

A common method for estimating manifold dimension is *principal component analysis* (PCA) [30], which derives a projected subspace where the data have maximum variance. Specifically, given a centered data set $\mathcal{X} = \{x_i \in \mathbb{R}^D\}_{i=1}^N$ and its corresponding matrix $X \in \mathbb{R}^{N\times D}$, we compute its covariance matrix $C = \frac{1}{N-1}X^\top X$ and perform eigen-decomposition as $C = \Gamma D \Gamma^\top$, where $D = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ and $\Gamma = [v_1, v_2, \ldots, v_N]$. λ_i is the i-th largest non-negative eigenvalue of C with v_i the corresponding eigenvector. For any variable x, the i-th principal component is given by $y_i = v_i x$. According to the definition, we have $\text{var}(y_i) = \lambda_i$, where $\text{var}(\cdot)$ denotes the variance operation. Then, a common criterion to estimate dimension is to find the minimum d that satisfies:

$$\frac{\sum_{i=1}^d \text{var}(y_i)}{\sum_{i=1}^N \text{var}(y_i)} = \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^N \lambda_i} > \beta, \quad 0 < \beta < 1. \tag{2}$$

When β gets close to 1, the above Eq. (2) essentially involves counting the number of large eigenvalues and omitting small ones. Since the squared singular values of X are proportional to the corresponding eigenvalues of C, this estimation is equivalent to the *effective rank* (ER) of X that treats its small singular values as zeros. Based on this finding, we employ the ER definition in [47]:

$$\text{ER}(X) = \exp\left(-\sum_{i=1}^N p_i \log p_i\right), \tag{3}$$

where p_i denotes the l_1-normalized singular value. Based on ER, we further introduce *local effective rank*, which is related to the modeling scale of motion unit defined formally in Sec. IV-A. Specifically, when applying ER to measuring local intrinsic dimension for skeleton sequences, we compute it on the set $\mathcal{X}$ composed of motion unit-scale representations extracted from a certain temporally cropped sequence, and term such a measure as local effective rank (LER). As derived from PCA, LER is more suitable for evaluating discrete representations than fractal dimension measures.

B. Local Dimension Collapse Analysis With LER

In this part, we analyze LER's properties as a rank estimator to indicate local dimension collapse, and further reveal the negative impact of local dimension collapse on TAS with LER.

1) LER for Rank Estimation: We first give fundamental properties linking LER to the rank of local representation matrices, demonstrating the efficacy of LER as a rank estimator. We denote $Z \in \mathbb{R}^{d\times m}$ as the matrix formed by a set of local representation vectors, with m representing the number of vectors and d denoting the dimensionality of each representation vector.

Proposition 1: The rank of local representations is a tight upper bound for LER, satisfying:

$$1 \le \text{LER}(Z) \le \text{rank}(Z) \le n, \tag{4}$$

where $\text{rank}(\cdot)$ denotes the rank operator and $n = \min\{m, d\}$. The equality holds if and only if the non-zero singular values are uniformly distributed, *i.e.*,

$$\sigma_i = \sigma_j \neq 0, \ \forall i, j \le \text{rank}(Z), \tag{5}$$

where σ_i denotes the i-th singular value.

Furthermore, in practice, we typically treat small singular values as zero to estimate a lower rank r for the matrix, which can significantly compress the representation matrix within a small reconstruction error threshold by storing fewer singular values and their corresponding singular vectors. This strategy is termed low-rank approximation, which can effectively mitigate the impact of noise and more robustly estimate the information content of the matrix. Following this, we propose to experimentally examine the relation between the minimal achievable rank r and LER when performing low-rank approximation.

Specifically, we conduct random sampling on skeleton representations encoded by a vanilla ViT [48] fine-tuned from scratch on PKUMMD I dataset [49] at different training epochs. For each sampled representation matrix, we compute both its LER value and the minimal rank r achievable under different error thresholds. The error is defined as the Frobenius norm of the error matrix, i.e., the sum of squared element-wise difference between the original matrix and the reconstructed matrix. To ensure that the approximated low-rank representations preserve the essential semantic information of the original ones, a strict error threshold is necessary. We selected 0.1, 0.5, and 2.0 because they are negligible (*e.g.*, <0.2%) compared to the magnitude of the representation matrix values, whose L1 norm is about 1,900 in average. The experimental results in Figure 2a demonstrate positive linear correlations between the LER values and the minimal r values. Formally, we compute the Pearson correlation coefficient (PCC) that measures the linear relationship between two variables, ranging from -1 to 1, to validate this correlation, and obtain high values of over 0.99 under all thresholds.

Moreover, LER possesses two invariance properties, which are consistent with matrix rank:

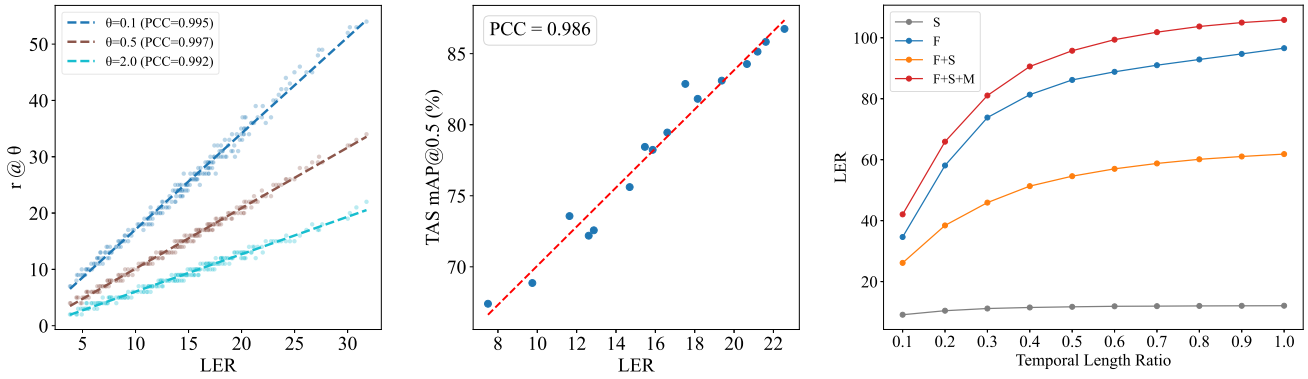
Proposition 2 (Scale Invariance): For all $c \neq 0$, the following equality holds:

$$\text{LER}(c \cdot Z) = \text{LER}(Z). \tag{6}$$

Proposition 3 (Orthogonal Invariance): For any orthogonal matrix $U \in \mathbb{R}^{d\times d}$, the following equality holds:

$$\text{LER}(UZ) = \text{LER}(Z). \tag{7}$$

Owing to the properties, LER effectively estimates the rank of local representation matrices and indicates the extent of local dimension collapse.



(a) LER vs. minimal r. (b) LER vs. TAS performance. (c) LER for learning at different temporal scales.

Fig. 2. The visualization results of local dimension collapse analysis based on LER. (a) shows the quasi-linear relation (PCC >0.99) between LER and the minimal achievable r under low-rank approximation. (b) shows the positive correlation (PCC = 0.986) between the LER and downstream TAS performance on the PKUMMD I dataset. (c) demonstrates the LER for the models with learning objectives of different scales, computed within local regions of varying temporal lengths. The length ratios are relative to the frame number of each cropped sequence, *i.e.*, 120. F, S, and M denote models trained with frame-, sequence- and motion unit-scale objectives, respectively.

2) Rank for Information Measurement: Given the nice properties of LER positively correlated with rank, we are motivated to explore the intrinsic relationship between the rank and the information content of representation matrices, which further reveals the impact of local dimension collapse. Specifically, we employ the coding length of the local representation matrix as the measurement for information content and use low-rank approximation to compress the matrix.

Lemma 1: Denote $Z \in \mathbb{R}^{d\times m}$ as the matrix formed by a set of local representation vectors, with m representing the number of vectors and d denoting the dimensionality of each representation vector. Assume each real value has a fixed coding length of b bits, and the coding length L_r under the optimal rank-r approximation of matrix Z is given by:

$$L_r = b \cdot r \cdot (d + m + 1). \tag{8}$$

Lemma 1 reveals that the minimal coding length is proportional to the achievable minimal rank r of the representation matrix. Since LER is linearly positively correlated with r, it serves as an indicator of this coding length. This confirms that local dimension collapse, reflected by low LER values, degrades the minimal coding length of local representations and thus results in uninformative latent spaces. In the context of TAS, this loss of information leads to imprecise temporal boundary predictions, which we will further examine below.

3) LER for Indicating TAS Performance: As established in prior works [29], [45], [50], dimension collapse hinders downstream perception. When it occurs locally in action modeling, the subtle distinctions between consecutive frames become ambiguous, *i.e.*, temporally adjacent but semantically distinct motion patterns may be mapped to similar representations. Consequently, TAS tasks that require precise frame-wise classification will be significantly affected by the local dimension collapse problem, which can be revealed by the LER measure. We examine the relation between LER and downstream TAS performance. By fine-tuning a vanilla ViT as previously mentioned, we show the LER value and corresponding TAS precision at different training epochs in Figure 2b. The mean average precision improves as the LER increases, with a high PCC value of 0.986. It verifies the positive correlation between LER value and TAS precision, confirming the utility of LER for the subsequent analysis.

C. Analysis of Learning at Different Scales

In this part, we analyze skeleton-based action representation learning with objectives of different temporal modeling scales. With the LER measure, we reveal the local dimension collapse problem in existing sequence-scale learning and frame-scale learning methods, and propose a novel motion unit-scale learning mechanism to enhance local intrinsic dimension. The implementations of sequence- and motion unit-scale learnings follow BYOL [43], while frame-scale learning follows MAMP [15]. Training and testing are performed on the NTU-RGB+D 60 dataset [51]. Figure 2c shows the results, which will be analyzed in detail.

1) Sequence-Scale Learning Exhibits Severe Local Dimension Collapse: Sequence-scale learning does not directly constrain local short-term representations. Instead, it indiscriminately introduces corresponding sequence-scale information to them. This leads to similar representations for temporally adjacent motion units, resulting in significant local dimension collapse as shown in the gray curve in Figure 2c. Consider an extreme scenario where sequence-scale learning constrains the representation clusters to form a globally high-dimensional manifold, while intra-cluster representations collapse to their centroids. In this case, the LER value of the representation space degenerates to 1 regardless of the global intrinsic dimension. This reveals the critical limitation of sequence-scale learning without local diversity constraints.

2) Frame-Scale Learning Alleviates But Still Suffers From Local Dimension Collapse: Frame-scale learning discriminately optimizes each motion unit representation with frame-scale action details, which introduces fine-grained constraints that expand the local dimension as shown in the blue curve in Figure 2c. However, it focuses on local motion patterns within each sequence, lacking awareness of inter-sequence relations. As a result, the local intrinsic dimension of the

latent representations is bounded by the dimensionality of the corresponding original action sequence. Since skeleton data is highly compact, its dimensionality is low, often even lower than that of the corresponding representation vectors. This inevitably leads to local dimension collapse, making it difficult for the model to actively explore more diverse representations for richer semantics, such as sequence-scale semantics or potential motion patterns not present in training data. Moreover, fine-grained reconstruction tasks focus more on optimizing principal components of the latent space as demonstrated in work [52], which can degrade the local intrinsic dimension and produce uninformative representations for perception.

3) Combining Frame-Scale and Sequence-Scale Learning Suffers From Semantic Gap: To help frame-scale learning methods perceive more sequence-scale semantics, a straightforward idea is to combine it with sequence-scale learning [13], [41]. We investigate this kind of approach but find that its LER value decreases, as shown in the orange curve in Figure 2c. As analyzed before, sequence-scale learning can hardly help regularize the local structure of representations; Instead, frame-scale learning brings much semantically irrelevant details, causing inconsistency when directly combined with sequence-scale learning. This semantic gap leads to instability and insufficiency in optimization, thereby reducing the semantic information captured in the representations and lowering the local intrinsic dimension.

4) Motion Unit-Scale Learning Improves Local Intrinsic Dimension: Theoretically, we verify that the motion unit-scale learning can effectively alleviate local dimension collapse because of:

Proposition 4 (Lower Bound of LER): The lower bound of LER is inversely proportional to the squared Frobenius norm of the similarity matrix of local representations. Specifically, for a matrix $Z \in \mathbb{R}^{d\times m}$ composed of l_2-normalized representation vectors, the following inequality holds:

$$\mathrm{LER}(Z) \geq \frac{m^2}{\|Z^\top Z\|_F^2}, \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm. Equality holds when either all singular values are equal or there is only one non-zero singular value.

Motion unit-scale learning can enrich the local short-term representations to reduce the similarity among them, and thus minimize $\|Z^\top Z\|_F$. Therefore, Proposition 4 implies that motion unit-scale learning can increase the lower bound of LER to mitigate local dimension collapse. Besides, the motion unit scale acts as an essential bridge between the low-level, noisy details of frame-scale learning and the high-level, abstract semantics of sequence-scale learning. Specifically, the motion units distill information from meaningful short-term motion patterns, creating intermediate-scale representations that can be effectively aligned with global context, which mitigates the semantic gap and boosts the local intrinsic dimension as exhibited in the red curve of Figure 2c.

Through theoretical and empirical analysis, we demonstrate the importance of motion unit-scale learning and that a multi-scale learning paradigm can effectively expand a high-dimensional latent space locally. Inspired by the findings, we propose our novel Local Dimension Enhancement learning framework involving a motion unit-based multi-scale learning module in the following section.

IV. Local Dimension Enhancement Learning Framework

In this section, we propose our novel learning framework, Local Dimension Enhancement (LoDE). Its overall pipeline is shown in Figure 3. It integrates a multi-scale action semantics learning module and an LER regularization loss to enhance the local intrinsic dimension of the motion unit-scale representations for more effective skeleton-based TAS.

A. Masked Motion Unit Modeling

The framework embeds motion units to capture short-term dynamics. It then uses a Siamese Transformer encoder with masking to process both masked and unmasked views, learning spatio-temporally aware representations.

1) Motion Unit-Scale Skeleton Embeddings: To construct representations at the motion unit scale, we begin by performing temporal aggregation on the input skeleton sequence. Formally, given a skeleton sequence $X \in \mathbb{R}^{T_0\times V\times 3}$, we reshape it into a series of non-overlapping temporal clips, yielding $X' \in \mathbb{R}^{T\times V\times 3l}$. T_0, V, and l respectively represent the number of initial frames, the number of joints, and the temporal length of each motion unit, where $T = T_0/l$. Each vector of dimension $3l$ within X' constitutes a motion unit, which is designed to capture short-term motion semantics more effectively than individual frames while being more fine-grained than entire sequences. The motion units are separately embedded into C dimension tokens with a projection matrix: $\mathrm{Embed}(X') \in \mathbb{R}^{T\times V\times C}$. Then, learnable spatial and temporal positional embeddings are added. Finally, these tokens are flattened into $E \in \mathbb{R}^{TV\times C}$.

2) Masking Strategy: To learn robust and semantics-rich representations, we employ masked modeling on the embedded tokens. Specifically, a high masking ratio is crucial for forcing the model to infer missing spatial-temporal context rather than relying on local redundancies, which introduces more action semantics into representations and leads to a high-dimensional latent space. Formally, we employ random masking for the embedded tokens with a ratio of r, yielding $K = \lceil(1-r)\times T\times V\rceil$ visible tokens $E_m \in \mathbb{R}^{K\times C}$. In practice, we generate the mask $M \in \{0,1\}^{(T\times V)}$ with an extremely high masking ratio $r = 90\%$ following the works on videos [53], [54].

3) Siamese Transformer Encoder: We employ the vanilla Transformer as the encoder $\mathcal{E}(\cdot)$ to flexibly model the information interaction among different motion units, which consists of alternating blocks of multi-head self-attention and multi-layer perception. Using a Siamese modeling approach, the motion unit representations for both the masked and unmasked view are obtained by the encoders with the same weights, *i.e.*, $Z_m^{\mathrm{motion}} = \mathcal{E}(E_m)$ and $Z^{\mathrm{motion}} = \mathcal{E}(E)$.

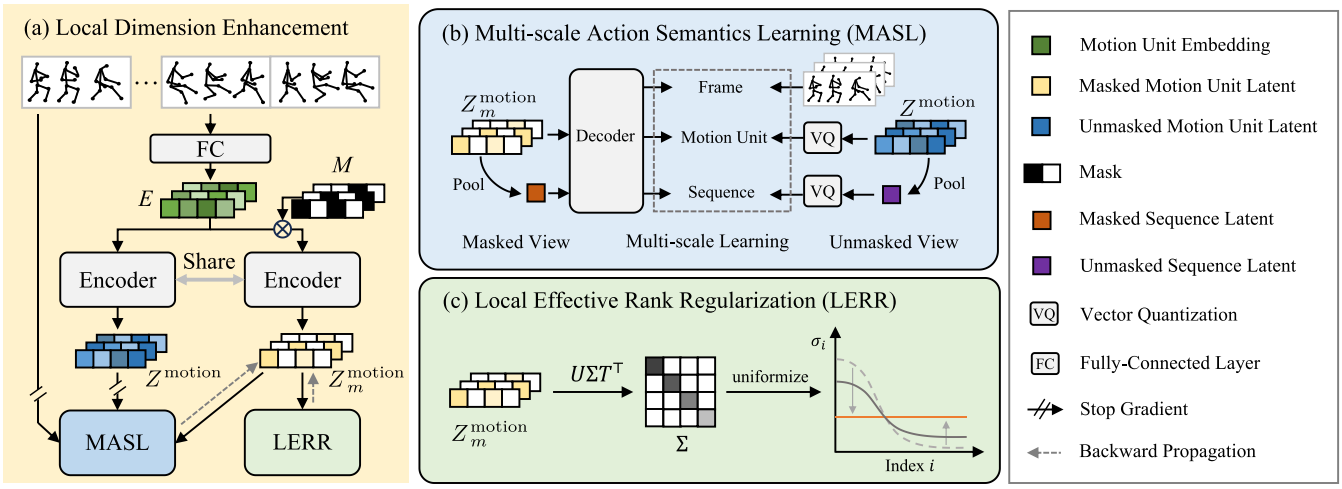


Fig. 3. Illustration of our Local Dimension Enhancement (LoDE) framework. (a) LoDE learns motion unit-scale representations by applying a masked modeling strategy with weight-sharing Siamese encoders. (b) The Multi-scale Action Semantics Learning (MASL) module captures multi-scale semantics by aligning the reconstructed objectives from the masked view with original skeletons and quantized representations from the unmasked view. (c) LER regularization (LERR) is applied to the masked view to encourage a more uniform singular value distribution and increase local intrinsic dimension.

B. Multi-Scale Action Semantics Learning (MASL)

In this module, the motion unit representations are refined to capture rich action semantics and form a high-dimensional latent space. We believe that such high-quality representations should simultaneously have three properties: 1) *multi-scale informative*, to be aware of long-range semantics and perceive short-scale transitions; 2) *diverse*, without compromising semantic integrity, different motion units have highly dissimilar representations; 3) *reconstructible*, the original data can be reconstructed with little information loss. Specifically, to achieve multi-scale awareness and diversity, we introduce a novel sequence-conditioned quantization modeling strategy that integrates the motion unit- and sequence-scale learning. To ensure that the representations remain reconstructible, we incorporate a frame-scale decoding objective. Details are introduced as follows.

1) Sequence-Conditioned Quantization Modeling: As analyzed in Sec. III-C, we perform motion unit-scale learning to diversify local representations and enhance LER. Meanwhile, we also integrate sequence-scale learning to perceive global action semantics. Instead of processing each scale separately, we introduce *sequence conditioning* strategy to mitigate the semantic gap with motion unit representations.

Specifically, the sequence-scale representation $Z_m^{\text{seq}} = \mathcal{P}_{\text{seq}}(\text{GAP}(Z_m^{\text{motion}}))$ is first obtained by a projector after pooling all visible motion unit representations, where GAP$(\cdot)$ represents global average pooling operation. Then we use it to pad the invisible positions in Z_m^{motion} and add positional embeddings to obtain the decoder input Z_{pad}. This introduces global semantics to motion unit representations.

To learn semantics-rich and diverse motion unit representations, we apply vector quantization (VQ) after the Siamese encoder for the target representations. Recent works have demonstrated VQ's efficacy in discriminative action understanding. Notably, SMQ [55] utilizes VQ to discretize continuous sequences into semantically rich "motion words", and HVQ [56] employs hierarchical codebooks to capture temporal dynamics of different levels. By mapping target representations to a discrete space, the VQ module filters out noise and enforces alignment on compact representations for motion patterns. In detail, a motion unit vector quantization module $\text{VQ}_{\text{motion}}(\cdot)$ is applied, which maintains a codebook and matches the input representation with the most similar vector in the codebook to obtain the target discrete representations $Y^{\text{motion}} = \text{VQ}_{\text{motion}}(Z^{\text{motion}})$. Similarly, we obtain the target sequence-scale tokens $Y^{\text{seq}} = \text{VQ}_{\text{seq}}(Z^{\text{seq}})$ using sequence vector quantization module $\text{VQ}_{\text{seq}}(\cdot)$. The codebooks are updated by the exponential moving average strategy [57], which is effective in preventing the collapse [43]. Then the sequence- and motion unit-scale objectives $\mathcal{L}_s$ and $\mathcal{L}_m$ are formulated by negative cosine similarity:

$$\begin{aligned} \mathcal{L}_m &= 1 - \cos(Y^{\text{motion}}, \mathcal{P}_{\text{motion}}(\hat{Z}_{\text{pad}})),\ \hat{Z}_{\text{pad}} = \mathcal{D}(Z_{\text{pad}}), \\ \mathcal{L}_s &= 1 - \cos(Y^{\text{seq}}, Z_m^{\text{seq}}), \end{aligned} \tag{10}$$

where $\mathcal{D}(\cdot)$ is the decoder and $\mathcal{P}_{\text{motion}}(\cdot)$ is a linear projector.

2) Frame-Scale Reconstructed Decoding: We adopt a Transformer decoder to predict original frame-scale motion patterns, which effectively preserves fine-grained information for latent representations. Technically, the projector $\mathcal{P}_{\text{frame}}(\cdot)$ is attached after $\mathcal{D}(\cdot)$ to obtain frame predictions, and *Mean Square Error* (MSE) loss is used for the frame-scale objective $\mathcal{L}_f$:

$$\mathcal{L}_f = \|\mathcal{P}_{\text{frame}}(\hat{Z}_{\text{pad}}) - \mathcal{T}(X)\|_2^2, \tag{11}$$

where $\mathcal{T}(\cdot)$ is a linear transformation to avoid learning shortcuts during reconstruction.

C. Local Effective Rank Regularization (LERR)

As validated in Sec. III-B, LER is an effective rank estimator and indicates the extent of local dimension collapse. Therefore, we propose to apply it as a direct regularization to the pre-training process. We perform the regularization on the motion unit representations Z_m^{motion} generated from the masked view.

Formally, we use the reciprocal of LER to encourage higher local intrinsic dimension for motion units:

$$\mathcal{L}_r = \frac{1}{\mathrm{LER}(Z_m^{\mathrm{motion}})}. \tag{12}$$

To address potential numerical instability during the back-propagation of SVD-based LER objective, specifically the risk of gradient explosion when singular values are close to zero, we employ a spectral shifting strategy, *i.e.*, a small regularization term ϵI (*e.g.*, $\epsilon = 10^{-6}$) is added to the representation matrix. This ensures that the singular values of the matrix remain strictly positive, thereby stabilizing the gradient computation and preventing numerical collapse during training.

The whole model is optimized by the MASL objectives with LERR loss. The overall loss is given by:

$$\mathcal{L} = \lambda_m \mathcal{L}_m + \lambda_s \mathcal{L}_s + \lambda_f \mathcal{L}_f + \lambda_r \mathcal{L}_r, \tag{13}$$

where λ_m, λ_s and λ_f are the weights for the motion unit-scale, sequence-scale and frame-scale objectives, respectively. λ_r is the weight for LER regularization.

V. Experiments

A. Datasets

To verify the effectiveness of the LoDE framework, we conduct experiments on the following datasets:

1) PKU Multi-Modality Dataset (PKUMMD) [31] is a large-scale dataset that covers long continuous sequences. Two phases are divided: PKUMMD I (PKU-I) and PKUMMD II (PKU-II). PKU-I is for large-margin action segmentation, while PKU-II is for small-margin segmentation. We adopt the cross-subject (Xsub) and the cross-view protocols (Xview) for both PKU-I and PKU-II.

2) Toyota Smarthome Untrimmed Dataset (TSU) [32] is a real-world dataset for TAS in daily living settings. It contains long-term composite activities where up to five actions can happen simultaneously in a single frame. Only 2D skeleton data is used, following LAC [21] for our experiments.

3) Bodies, Action and Behavior with English Labels Dataset (BABEL) [33] is a frame-scale annotated dataset containing about 43 hours of motion sequences and over 250 action categories from AMASS [58]. We follow BID [23] that creates three subsets of BABEL, pretrains the model on all the subsets, and applies fine-tuning separately.

4) NTU RGB+D Dataset 60 (NTU 60) [51] is a large-scale trimmed dataset containing 56,578 samples with 60 action categories and 25 joints. We employ it for self-supervised pre-training and study the transfer-learning performance of our method on PKUMMD.

5) Posetics Dataset (Posetics) [59] is a real-world action dataset that contains about 142,000 trimmed skeleton sequences with 13 body joints of 2D and 3D representations. It is also used for self-supervised pre-training and verifying the transfer learning performance on PKUMMD and TSU.

B. Implementation Details

1) Pre-Training Settings: The input sequence of 300 frames is cropped and interpolated to 120 frames, and the temporal size of each motion unit is $l = 4$. A vanilla ViT [48] is adopted as the backbone, and the MAE-based [40] architecture is built with an 8-layer encoder and a 3-layer decoder. The hidden dimension is 256 in attention layers and 1024 in feed-forward layers. We pre-train the models with a total batch size of 64 for 400 epochs. The base learning rate is set to 1e-3 and is reduced to 5e-4 gradually with cosine decay. The AdamW optimizer is employed with the momentum of $\beta_1, \beta_2 = 0.9, 0.95$ and the weight decay of 0.05. λ_m, λ_s, and λ_f are set to 0.1, 1.0, 1.0. λ_r is 1e-4 for NTU 60 and 1e-6 for Posetics and BABEL.

2) Downstream Evaluation Settings: For downstream evaluation, we apply *linear* and *fine-tuning* protocols. In linear evaluation, the pre-trained backbone is frozen, and a linear classifier is attached. The classifier is trained supervised by ground-truth labels for 100 epochs with a batch size of 128 and a learning rate decreasing from 0.01 to 0. In the fine-tuning evaluation, we attach an MLP head and fine-tune the whole network for 100 epochs with a batch size of 32. The learning rate increases linearly to 3e-4 from 0 in the first five warm-up epochs and then decreases to 1e-5 with cosine decay.

We evaluate on three untrimmed TAS datasets PKUMMD, TSU and BABEL. For PKUMMD, we pre-train on NTU 60 by default. Results of pre-training on Posetics are also highlighted in linear evaluation. Event-based mean Average Precision (mAP) is reported at different temporal Intersection over Union (tIoU) thresholds between the predicted and the ground truth intervals to evaluate the performance. For TSU, we pre-train the model with sequences from Posetics, and employ per-frame mean Average Precision (per-frame mAP) following LAC [21]. For BABEL, we follow the pretraining and evaluation protocols in BID [23]. A sliding window strategy is adopted to process long sequences following PCM [3] [13].

C. Fine-Tuning Evaluation Results

To validate the superiority and generalizability of our method, we conduct extensive experiments on different skeleton-based action datasets with varying difficulty levels. We compare the fine-tuning results of LoDE to other skeleton-based TAS methods, and some supervised learning and RGB-based methods are reported for reference.

1) Results on PKU-I: PKU-I is a laboratory-collected long-action sequence dataset with clear action intervals. It primarily evaluates models' capabilities in action semantic extraction and label prediction consistency. As shown in Table I, LoDE achieves remarkable segmentation performance compared to other TAS methods. Through multi-scale learning centered on motion units, LoDE learns a high-dimensional latent space with the information of both local temporal boundaries and global action semantics. Therefore, compared to frame-scale learning methods like MAMP and LAC, LoDE demonstrates more pronounced performance advantages at high tIoU thresholds, indicating better temporal consistency.

2) Results on PKU-II: Compared to PKU-I, PKU-II features significantly shorter action intervals, which demands models to possess stronger temporally fine-grained discriminative capability. Moreover, due to the limited scale of PKU-II, early TAS methods that directly fit training set labels are prone to overfitting. Nevertheless, as shown in Table II, LoDE

TABLE I

FINE-TUNING TAS PERFORMANCE ON PKU-I. MAP AT TIOU $\theta = 0.1$, $\theta = 0.3$ AND $\theta = 0.5$ ARE REPORTED. BOLD AND UNDERLINED VALUES DENOTE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY

Methods		Publication	Modality	Xsub mAP@θ (%)			Xview mAP@θ (%)		
				0.1	0.3	0.5	0.1	0.3	0.5
GRU-GD	[37]	ECCV 2018	RGB	88.0	86.8	80.1	-	-	-
SSTCN-GD	[60]	ICCV 2021	RGB	83.7	82.1	76.5	-	-	-
Augmented-RGB	[60]	ICCV 2021	RGB	86.3	84.5	81.1	-	-	-
Supervised Methods									
GRU-GD	[37]	ECCV 2018	Skeleton	85.7	84.6	78.4	-	-	-
TAP-B-M	[2]	TIP 2018	Skeleton	51.3	48.0	39.5	63.2	58.5	48.6
Beyond Joints	[61]	TIP 2018	Skeleton	87.4	-	81.1	95.3	-	91.1
Convolution Skeleton	[31]	TOMM 2020	Skeleton	49.3	31.8	12.1	54.9	36.6	16.3
Cui *et al.*	[62]	IET CV 2020	Skeleton	-	-	83.5	-	-	93.3
Self-Supervised Methods									
MAMP	[15]	ICCV 2023	Skeleton	91.3	91.0	88.1	95.6	95.5	93.8
LAC	[21]	ICCV 2023	Skeleton	91.8	90.2	88.5	93.9	-	-
PCM^3	[13]	ACM MM 2023	Skeleton	78.1	76.7	70.1	83.7	82.1	75.0
MacDiff	[17]	ECCV 2024	Skeleton	93.4	93.1	91.2	96.0	95.4	93.4
USDRL	[18]	TPAMI 2025	Skeleton	90.8	89.7	86.8	92.7	92.0	88.8
LoDE (Ours)		—	Skeleton	**94.1**	**93.7**	**92.2**	**96.6**	**96.3**	**94.4**

TABLE II

FINE-TUNING TAS PERFORMANCE ON PKU-II. MAP AT TIOU $\theta = 0.5$ IS REPORTED

Methods		Publication	Modality	mAP@0.5 (%)	
				Xsub	Xview
JCRRNN	[34]	ECCV 2016	RGB	5.9	4.6
UntrimmedNet	[63]	CVPR 2017	RGB	1.9	1.7
SW-LSTM	[31]	TOMM 2020	RGB	6.8	6.9
SW-BLSTM	[31]	TOMM 2020	RGB	8.3	6.8
SW-TPN	[64]	BMVC 2017	Skeleton	3.3	1.5
SW-STA-LSTM	[2]	TIP 2018	Skeleton	2.2	2.2
BLSTM	[31]	TOMM 2020	Skeleton	3.7	4.0
MAMP	[15]	ICCV 2023	Skeleton	23.6	10.3
PCM^3	[13]	MM 2023	Skeleton	17.3	5.8
MacDiff	[17]	ECCV 2024	Skeleton	27.1	19.0
USDRL	[18]	TPAMI 2025	Skeleton	20.5	6.2
LoDE (Ours)		—	Skeleton	**29.0**	**19.8**

TABLE III

FINE-TUNING TAS PERFORMANCE ON TSU. PER-FRAME MAP IS REPORTED

Methods		Publication	Modality	mAP (%)
TGM	[65]	ICML 2019	RGB	26.7
PDAN	[66]	WACV 2021	RGB	32.7
SD-TCN	[32]	TPAMI 2022	RGB	29.2
MS-TCT	[35]	CVPR 2022	RGB	33.7
Bi-LSTM	[67]	NN 2005	Skeleton	17.0
TGM	[65]	ICML 2019	Skeleton	26.7
SD-TCN	[32]	TPAMI 2022	Skeleton	26.2
LAC	[21]	ICCV 2023	Skeleton	34.1
SCS	[24]	AAAI 2025	Skeleton	35.1
LoDE (Ours)		—	Skeleton	**35.7**

TABLE IV

FINE-TUNING TAS PERFORMANCE ON THREE SUBSETS OF BABEL. MAP AT TIOU $\theta = 0.5$ IS REPORTED

Methods		Publication	mAP@0.5 (%)		
			Set-1	Set-2	Set-3
CoLA	[68]	CVPR 2021	2.15	10.32	8.99
FAC-Net	[69]	CVPR 2021	2.62	6.53	8.90
UP-TAL	[70]	CVPR 2022	32.37	30.49	25.19
LART	[71]	CVPR 2023	23.82	15.43	16.38
S-WTAL	[72]	AAAI 2023	21.70	19.55	20.36
BID	[23]	arXiv 2024	35.01	35.13	24.60
LoDE (Ours)		—	**41.45**	**48.45**	**48.31**

achieves significant performance improvements on PKU-II by introducing abundant temporally fine-grained action semantics through frame- and motion unit-scale learning. Besides, with LERR, the learned local latent space is spanned, which forces LoDE to capture more inherent motion semantics from the limited training data and alleviates the overfitting problem.

3) Results on TSU: TSU is an indoor-captured dataset characterized by concurrent actions, where each frame may correspond to multiple action labels. This necessitates frame-scale representations with richer spatial information to support multi-action classification. Benefiting from fine-grained reconstruction modeling and high local intrinsic dimension, LoDE encodes action representations with strong temporal discriminability while maintaining spatial information richness, thereby better supporting concurrent action segmentation tasks and achieving higher per-frame mAP in Table III.

4) Results on BABEL Subsets: The BABEL dataset, derived from AMASS that unifies 15 human motion datasets, exhibits more complex and diverse motion patterns compared to indoor-captured data. Compared to previous methods, LoDE learns a higher-dimensional latent space at the motion unit scale. Therefore, it captures more meaningful semantics for diverse motion patterns and achieves significant improvements on all three subsets, illustrating greater generalization capability and robustness, which is demonstrated in Table IV.

D. Linear Evaluation Results

We evaluate the TAS performance of LoDE on PKU-I using linear evaluation protocols as shown in Table V. Since LAC is pre-trained on Posetics and other compared methods are pre-trained on NTU 60, we pre-train our model on both datasets for fair comparisons. Our method achieves

TABLE V
TRANSFER-LEARNING RESULTS BY LINEAR EVALUATION ON PKU-I. MAP AT TIOU $\theta = 0.1$ AND $\theta = 0.5$ ARE REPORTED

Methods		Source Dataset	Pre-training	PKU-I Xsub@θ (%) 0.1	0.5	PKU-I Xview@θ (%) 0.1	0.5
LAC	[21]	Posetics	*w/o* labels	55.2	-	58.8	-
LAC-supervised	[21]	Posetics	*w* labels	61.8	-	62.4	-
LoDE (Ours)		Posetics	*w/o* labels	77.7	71.8	72.4	66.0
SCS	[24]	NTU 60	*w* labels	80.9	76.4	-	-
MAMP	[15]	NTU 60	*w/o* labels	81.8	77.1	85.4	80.8
MacDiff	[17]	NTU 60	*w/o* labels	87.0	80.0	93.1	89.5
LoDE (Ours)		NTU 60	*w/o* labels	**87.6**	**83.1**	**93.5**	**90.1**

TABLE VI
TRANSFER-LEARNING RESULTS BY FINE-TUNING EVALUATION ON ACTION RECOGNITION

Methods		Publication	Accuracy (%) NTU	PKU-I	PKU-II
AimCLR	[10]	AAAI 2022	86.9	-	51.6
SkeletonMAE	[14]	ICCV 2023	88.5	-	58.4
MAMP	[15]	ICCV 2023	**93.1**	96.3	70.6
MacDiff	[17]	ECCV 2024	92.7	96.5	**72.2**
Colorization	[73]	TPAMI 2024	89.1	-	58.1
I^2MD	[19]	IJCV 2025	86.5	-	60.7
HSARL	[20]	CVPR 2025	-	-	64.3
LoDE (Ours)		—	92.4	**97.0**	71.9

TABLE VII
ABLATION STUDY ON MULTI-SCALE OBJECTIVES

Scales	mAP@0.5 (%)
F	77.1
F + M	79.0
F + S	76.3
F + M + S	**80.9**

TABLE VIII
ABLATION STUDY ON LOCAL DIMENSION REGULARIZER

Measures	mAP@0.5 (%)
w/o Reg.	80.9
MOM	80.2
LER	**83.1**

significant improvements, verifying the high quality of the learned representations. Note that our model is pre-trained in a self-supervised manner, *i.e.*, without using labels on source datasets, but it still surpasses supervisedly pre-trained LAC and SCS. This implies that our model can effectively capture high-level action semantics from the well-designed multi-scale learning task without explicit semantic labels.

It is also notable that there are significant differences in the skeleton formats between Posetics and PKU-I. Therefore, transfer learning between them is much more challenging compared to transferring from NTU 60 to PKU-I. Nevertheless, benefiting from the informative representations with high local intrinsic dimension, LoDE performs well under this setting, *e.g.*, achieving mAP@0.1 22.5% higher than LAC under the Xsub protocol, which highlights its strong transfer ability.

E. Action Recognition Results

Although LoDE is designed for TAS, it shows competitive performance on action recognition. As shown in Table VI, we pre-train on NTU 60 and fine-tune on the Xsub benchmark of NTU 60, PKU-I (trimmed) and PKU-II (trimmed), compared with well-performed skeleton-based action recognition methods. LoDE achieves comparable recognition accuracy to prior recognition methods, demonstrating excellent compatibility with the action recognition task.

F. Ablation Study

We carry out a series of ablation experiments on the PKU-I cross-subject benchmark, with self-supervised pre-training on NTU 60 and the linear evaluation protocol.

1) Effectiveness of Multi-Scale Objectives: Table VII shows the results with different scales of objectives in MASL without LERR. F, S, and M respectively represent adopting frame-, sequence- and motion unit-scale learning. Compared to frame-scale learning, refining it with sequence-scale learning encounters slight performance degradation, implying a semantic gap. In contrast, combining frame- and motion unit-scale learning improves the performance. When we adopt all the objectives, the performance is further boosted, verifying the importance of all our MASL objectives.

2) Choice of Local Dimension Regularizer: We adopt the model without LERR as the baseline, and test the performance when applying LER and a fractal dimension measure MOM [29] as regularization, respectively. As shown in Table VIII, the TAS performance degrades slightly when MOM is added as the local dimension constraint due to its instability. In contrast, applying the stable manifold dimension measure LER improves the performance.

3) Sensitivity to LERR Weight: Figure 4 illustrates the model's sensitivity to the LERR weight. Optimal performance is achieved at a weight of 1e-4. As the weight decreases, the regularization effect diminishes, and performance approaches the baseline. Conversely, an excessively large weight severely hampers performance. This is because LERR's primary role is to structure the latent space, not to encode action semantics directly. An overly strong weight disrupts the learned semantic relationships, underscoring the need for a moderate value.

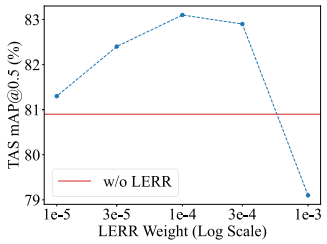


Fig. 4. Ablation study on the weight for LERR.

TABLE IX

ABLATION STUDY ON DESIGNS FOR INTEGRATING MULTI-SCALE LEARNING. VQ AND SC REPRESENT VECTOR QUANTIZATION AND SEQUENCE CONDITIONING, RESPECTIVELY

Modules		mAP@θ (%)		
VQ	SC	0.1	0.3	0.5
		74.9	73.7	70.2
✓		81.7	81.2	78.1
	✓	69.7	67.7	60.9
✓	✓	**86.5**	**85.5**	**80.9**

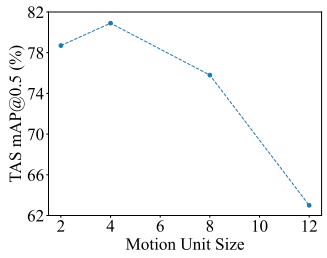


Fig. 5. Ablation study on the temporal size of motion units.

4) Designs for Integrating Multi-Scale Learning: We verify the effects of our MASL designs to integrate different scales in Table IX. The baseline model omits vector quantization and sequence conditioning. We observe that the TAS performance improves when both the designs are adopted. Also note that without vector quantization, sequence conditioning tends to overfit to semantically irrelevant information and leads to lower performance. It implies that they are not independent components but a tightly integrated mechanism to learn diverse, multi-scale representations.

5) Impact of Motion Unit Size: We first compare the effects of different motion unit sizes without LERR. As shown in Figure 5, the TAS performance peaks at size 4. This is intuitive: smaller units fail to capture sufficient temporal context, while larger units lose the fine-grained resolution necessary for dense predictions in TAS, confirming the trade-off between semantic richness and temporal precision. We further analyze the sensitivity of the motion unit size on PKU-I Xsub, TSU and BABEL as shown in Table X. While the optimal motion unit size varies slightly due to dataset-specific temporal characteristics, our method generally achieves competitive performance compared to state-of-the-art methods. This confirms that our framework is robust to variations in motion unit length and maintains superior performance without requiring strict hyperparameter tuning.

TABLE X

SENSITIVITY ANALYSIS ON MOTION UNIT SIZE (l) ON PKU-I XSUB, TSU, AND BABEL. WE REPORT AVERAGE RESULTS OF THE THREE SUBSETS ON BABEL. SOTA METHODS REFER TO MACDIFF [17] ON PKU-I XSUB, SCS [24] ON TSU, AND BID [23] ON BABEL, RESPECTIVELY

Methods	PKU-I Xsub mAP@0.5 (%)	TSU mAP (%)	BABEL mAP@0.5 (%)
SoTA	80.0	35.1	31.6
LoDE ($l = 2$)	81.8	35.4	**46.1**
LoDE ($l = 4$)	**83.1**	**35.7**	44.1
LoDE ($l = 6$)	80.8	34.7	39.6

TABLE XI

ABLATION STUDY ON STUDENT-TEACHER NETWORK ARCHITECTURE WITH MOMENTUM

Architecture	Momentum	mAP@0.5 (%)
Siamese	–	83.1
EMA	0.999	77.5
	0.990	80.5

TABLE XII

COMPARISON OF PRE-TRAINING EFFICIENCY AND PERFORMANCE. "S" REPRESENTS SEQUENCE-SCALE LEARNING

Methods	Train Speed (samples/s) ↑	Total Time (hours) ↓	mAP@0.5 (%) ↑
MAMP [15]	167.83	14.6	77.1
LoDE *w/o* LERR	196.83	10.1	80.9
LoDE *w/o* S	128.39	14.3	81.8
LoDE	126.62	11.4	**83.1**

6) Utilization of Siamese Architecture: We investigate the impact of the teacher network update strategy. While Exponential Moving Average (EMA) is commonly used in contrastive learning [39], [43], we argue that it introduces an excessively large distributional gap which hinders effective representation alignment, since we have applied a high mask ratio (90%) and VQ modules to avoid learning short-cut. To verify this, we compare our Siamese architecture with EMA-based teachers. As shown in Table XI, the Siamese approach yields superior performance, confirming that immediate weight synchronization is a better choice for our proposed method.

7) Pre-Training Efficiency-Performance Trade-off: We evaluate the computational cost of the LERR and sequence-scale learning with default hyperparameters applied for NTU pre-training. We note that the total training time may fluctuate due to I/O latency and data loading overhead. Therefore, training throughput (samples/s) serves as a more robust and reliable metric for assessing computational cost. As shown in Table XII, including LER reduces pre-training throughput by approximately 1.6x (from 196.83 to 126.62 samples/s), and 1.3x compared to MAMP [15]. On the other hand, sequence-scale learning has a negligible impact on training efficiency. Notably, the overhead is strictly limited to the pre-training phase and does not affect fine-tuning or inference latency. Given the significant performance improvement, we consider the pre-training cost of full model an acceptable trade-off for enhanced representation quality.

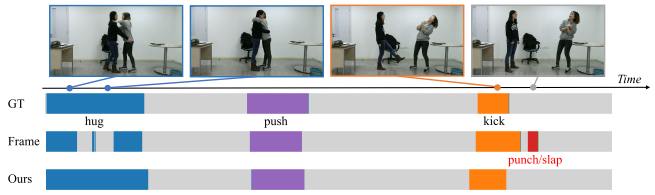


Fig. 6. TAS results of frame 300 to 1100 from "0299-M" in PKU-I.

G. Visualization Results

To qualitatively demonstrate the superior performance of our method, we present the segmentation result of sample "0299-M" from the validation set of PKU-I in Figure 6 with linear evaluation. More visualization results can be found in the Supplement. The result obtained by the model pre-trained with frame-scale learning is also shown for comparison, which is marked as "Frame". The results of Frame exhibit two notable segmentation errors in the example. First, during the relatively static phases in the middle of the "hug" action, the Frame model misjudges them as no-action segments. Second, after the "kick" action is completed, the Frame model misidentifies the following no-action segment as "punch/slap" due to the person being kicked appearing to be hit. Due to the local dimension collapse, Frame model fails to capture meaningful sequence-scale semantics and tends to generate predictions that are inconsistent with the context in these confusable scenarios. In contrast, our method achieves higher segmentation quality with more accurate temporal action boundaries and better temporal coherence, demonstrating its advantages in both fine-grained perception and long-range consistency.

VI. Conclusion

We propose the Local Dimension Enhancement learning framework to mitigate local dimension collapse in skeleton-based TAS. By introducing a manifold dimension measure LER and an intermediate modeling scale motion unit, we quantify local dimension collapse and analyze this problem in SSL at different scales. Based on the analysis, we propose MASL to enhance representation diversity by integrating sequence-, frame-, and motion unit-scale learning, while LERR further improves local intrinsic dimension. Theoretical analysis and experimental results on PKUMMD, TSU, and BABEL datasets demonstrate the effectiveness of the LoDE learning framework.

References

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7444–7452.

[2] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.

[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, Jun. 2019, pp. 12018–12027.

[4] 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. (ICCV)*, Oct. 2021, pp. 13339–13348.

[5] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 180–189.

[6] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[7] 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.

[8] 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.

[9] Z. Wu et al., "SelfGCN: Graph convolution network with self-attention for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 4391–4403, 2024.

[10] 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 *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 762–770.

[11] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 3427–3435.

[12] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2363–2372.

[13] J. Zhang, L. Lin, and J. Liu, "Prompted contrast with masked motion modeling: Towards versatile 3D action representation learning," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7175–7183.

[14] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, and L. Lin, "SkeletonMAE: Graph-based masked autoencoder for skeleton sequence pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 5583–5595.

[15] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li, "Masked motion predictors are strong 3D action representation learners," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10147–10157.

[16] L. Lin, J. Zhang, and J. Liu, "Mutual information driven equivariant contrastive learning for 3D action representation learning," *IEEE Trans. Image Process.*, vol. 33, pp. 1883–1897, 2024.

[17] L. Wu, L. Lin, J. Zhang, Y. Ma, and J. Liu, "MacDiff: Unified skeleton modeling with masked conditional diffusion," in *Proc. European Conf. Comput. Vis.*, Milano, Italy, 2024, pp. 110–128.

[18] H. Wang et al., "Foundation model for skeleton-based human action understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 1, pp. 1–16, Jan. 2026.

[19] Y. Mao, J. Deng, W. Zhou, Z. Lu, W. Ouyang, and H. Li, "I^{2}MD: 3D action representation learning with inter{-} and intra-modal mutual distillation," *Int. J. Comput. Vis.*, vol. 133, no. 7, pp. 4944–4961, Jul. 2025.

[20] H. Wang, X. Ma, J. Kuang, and J. Gui, "Heterogeneous skeleton-based action representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 19154–19164.

[21] D. Yang et al., "LAC–latent action composition for skeleton-based action segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 13633–13644.

[22] Y. Chen et al., "Hierarchically self-supervised transformer for human skeleton representation learning," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel: Springer, Oct. 2022, pp. 185–202.

[23] Q. Fang, C. Tang, S. Ma, and Y. Yang, "BID: Boundary-interior decoding for unsupervised temporal action localization pre-training," 2024, *arXiv:2403.07354*.

[24] H. Tian and P. Payeur, "Stitch, contrast, and segment: Learning a human action segmentation model using trimmed skeleton videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 7365–7373.

[25] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 2644–2651.

[26] X. Fang, J. Li, Q. Sun, and B. Wang, "Rethinking the uniformity metric in self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Vienna, Austria, 2024, pp. 1–22.

[27] X. Weng et al., "Modulate your spectrum in self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Vienna, Austria, 2023, pp. 1–28.

[28] L. Fournier, A. Patel, M. Eickenberg, E. Oyallon, and E. Belilovsky, "Preventing dimensional collapse in contrastive local learning with subsampling," in *Proc. Int. Conf. Mach. Learn. Workshops*, 2023, pp. 1–8.

[29] H. Huang, R. J. G. B. Campello, S. Erfani, X. Ma, M. E. Houle, and J. Bailey, "LDReg: Local dimensionality regularized self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Vienna, Austria, 2024, pp. 1–26.

[30] M. Fan, N. Gu, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of data by principal component analysis," 2010, *arXiv:1002.2050*.

[31] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–24, May 2020.

[32] R. Dai et al., "Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2533–2550, Feb. 2023.

[33] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "BABEL: Bodies, action and behavior with English labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 722–731.

[34] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 203–220.

[35] R. Dai, S. Das, K. Kahatapitiya, M. S. Ryoo, and F. Brémond, "MS-TCT: Multi-scale temporal ConvTransformer for action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20041–20051.

[36] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10105–10115.

[37] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 174–192.

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

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

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

[42] T. Chen, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 2024, pp. 1597–1607.

[43] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[44] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[45] J. Li, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, Lisbon, Portugal, 2021, pp. 1–17.

[46] M. E. Houle, "Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications," in *Proc. Int. Conf. Similarity Search Appl.*, Munich, Germany, 2017, pp. 64–79.

[47] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *Proc. 15th Eur. Signal Process. Conf.*, Poznan, Poland, Sep. 2007, pp. 606–610.

[48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[49] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," in *Proc. ACM Int. Conf. Multimedia Workshop*, 2017, pp. 1–8.

[50] Q. Garrido, R. Balestriero, L. Najman, and Y. LeCun, "RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10929–10974.

[51] 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. (CVPR)*, Jun. 2016, pp. 1010–1019.

[52] R. Balestriero and Y. LeCun, "Learning by reconstruction produces uninformative features for perception," 2024, *arXiv:2402.11337*.

[53] Y. Song, Z. Tong, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst. 35*, 2022, pp. 10078–10093.

[54] H. Fan, C. Feichtenhofer, K. He, and Y. Li, "Masked autoencoders as spatiotemporal learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 35946–35958.

[55] U. Gökay, F. Spurio, D. R. Bach, and J. Gall, "Skeleton motion words for unsupervised skeleton-based temporal action segmentation," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2025, pp. 12101–12111.

[56] F. Spurio, E. Bahrami, G. Francesca, and J. Gall, "Hierarchical vector quantization for unsupervised action segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, France, 2025, pp. 6996–7005.

[57] A. Razavi, A. V. D. Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[58] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5441–5450.

[59] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond, "UNIK: A unified framework for real-world skeleton-based action recognition," in *Proc. Brit. Mach. Vis. Conf.*, France, 2021, pp. 1–13.

[60] R. Dai, S. Das, and F. Bremond, "Learning an augmented RGB representation with cross-modal knowledge distillation for action detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13033–13044.

[61] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4382–4394, Sep. 2018.

[62] R. Cui, A. Zhu, J. Wu, and G. Hua, "Skeleton-based attention-aware spatial–temporal model for action detection and recognition," *IET Comput. Vis.*, vol. 14, no. 5, pp. 177–184, Aug. 2020.

[63] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6402–6411.

[64] Y. Hu, C. Liu, Y. Li, and J. Liu, "Temporal perceptive network for skeleton-based action recognition," in *Proc. Procedings Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[65] A. Piergiovanni and M. Ryoo, "Temporal Gaussian mixture layer for videos," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5152–5161.

[66] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond, "PDAN: Pyramid dilated attention network for action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2970–2979.

[67] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.

[68] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "CoLA: Weakly-supervised temporal action localization with snippet contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16010–16019.

[69] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8002–8011.

[70] C. Zhang, T. Yang, J. Weng, M. Cao, J. Wang, and Y. Zou, "Unsupervised pre-training for temporal action localization tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14031–14041.

[71] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, and J. Malik, "On the benefits of 3D pose and tracking for human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mali, Jun. 2023, pp. 640–649.

[72] Q. Yu and K. Fujiwara, "Frame-level label refinement for skeleton-based weakly-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 3322–3330.

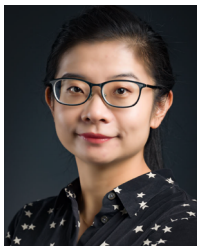
[73] S. Yang, J. Liu, S. Lu, E. M. Hwa, Y. Hu, and A. C. Kot, "Self-supervised 3D action representation learning with skeleton cloud colorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 509–524, Jan. 2024.

Shaofan Sun received the B.S. degree in computer science from Peking University, Beijing, China, in 2025, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His current research interests include human action understanding and representation learning.

Lilang Lin (Graduate Student Member, IEEE) received the B.S. degree in data science from Peking University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition, self-supervised learning, and unsupervised learning.

Jiahang Zhang received the B.S. degree in computer science from Peking University, Beijing, China, in 2023, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition and self-supervised learning.

Ling-Yu Duan (Member, IEEE) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, Australia, in 2008. He has been a Full Professor with the National Engineering Laboratory of Video Technology (NELVT), School of Computer Science, Peking University (PKU), China, and has served as the Associate Director of the Rapid-Rich Object Search Laboratory (ROSE), a joint laboratory between Nanyang Technological University (NTU), Singapore, and PKU, since 2012. He has been with the Peng Cheng Laboratory, Shenzhen, China, since 2019. He has published about 200 research articles. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics. He is a member of the MSA Technical Committee in the IEEE-CAS Society. He serves as the Area Chair of ACM MM and IEEE ICME. He received the IEEE ICME Best Paper Award in 2019/2020, the IEEE VCIP Best Paper Award in 2019, and EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, China Patent Award for Excellence in 2017, and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was the Co-Editor of the MPEG Compact Descriptor for Visual Search (CDVS) Standard (ISO/IEC 15938-13) and the MPEG Compact Descriptor for Video Analytics (CDVA) Standard (ISO/IEC 15938-15). He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Intelligent Systems and Technology*, and *ACM Transactions on Multimedia Computing, Communications, and Applications*.



Jiaying Liu (Fellow, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, 2010. She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia in 2015, supported by the Star Track Young Faculties Award. She is currently a Boya Distinguished Professor with the Wangxuan Institute of Computer Technology, Peking University, China. She has authored more than 100 technical articles in refereed journals and proceedings and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a Distinguished Member of CCF and a Senior Member of CSIG. She has been a member of the Multimedia Systems and Applications Technical Committee (MSA TC) and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in the IEEE Circuits and Systems Society. She has served as an ACM ICMR Steering Committee Member and the CAS Representative at the IEEE ICME Steering Committee. She was the APSIPA Distinguished Lecturer from 2016 to 2017. She was the General Chair of ACM MM Asia 2024 and the Technical Program Chair of ACM MM Asia 2025/2023/IEEE ICME 2021/ACM ICMR 2021/IEEE VCIP 2019. She has also served as the Deputy Editor-in-Chief for *APSIPA Transactions on Signal and Information Processing* and an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *IEEE Multimedia* magazine, and *Journal of Visual Communication and Image Representation*.