Foundation Model for Skeleton-Based Human Action Understanding

Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo-Sen Xie, Xin Geng, Senior Member, IEEE, and Liang Wang, Fellow, IEEE

Abstract—Human action understanding serves as a foundational pillar in the field of intelligent motion perception. Skeletons serve as a modality- and device-agnostic representation for human modeling, and skeleton-based action understanding has potential applications in humanoid robot control and interaction. However, existing works often lack the scalability and generalization required to handle diverse action understanding tasks. There is no skeleton foundation model that can be adapted to a wide range of action understanding tasks. This paper presents a Unified Skeleton-based Dense Representation Learning (USDRL) framework, which serves as a foundational model for skeleton-based human action understanding. USDRL consists of a Transformer-based Dense Spatio-Temporal Encoder (DSTE), Multi-Grained Feature Decorrelation (MG-FD), and Multi-Perspective Consistency Training (MPCT). The DSTE module adopts two parallel streams to learn temporal dynamic and spatial structure features. The MG-FD module collaboratively performs feature decorrelation across temporal, spatial, and instance domains to reduce dimensional redundancy and enhance information extraction. The MPCT module employs both multi-view and multi-modal self-supervised consistency training. The former enhances the learning of high-level semantics and mitigates the impact of low-level discrepancies, while the latter effectively facilitates the learning of informative multimodal features. We perform extensive experiments on 25 benchmarks across across 9 skeleton-based action understanding tasks, covering coarse prediction, dense prediction, and transferred prediction. Our approach significantly outperforms the current state-of-the-art methods. We hope that this work would broaden the scope of research in skeleton-based action understanding and encourage more attention to dense prediction tasks. This code is available at: https://github.com/wengwanjiang/FoundSkelModel.

Index Terms—Skeleton-Based Action Understanding, Human Foundation Model, Skeleton Foundation Model

1 Introduction

Skeleton-based human action understanding has tremendous applications in areas such as robotics, human-robot interaction, immersive virtual environments, assistive technology, rehabilitation and sports analytics. With advancements in pose estimation, skeleton-based action understanding is becoming an increasingly valuable tool in AI-driven motion perception. Compared to raw video or point clouds, skeleton data is more compact, lightweight, and computationally efficient, while also offering privacy-preserving advantages.

Supervised skeleton-based action recognition with deep learning has rapidly developed over the past decade. Representative methods include Hierarchical RNN [1], [2], Spatio-Temporal LSTM [3] Two-Stream RNN [4], Multi-Modal Representations [5], ST-GCN [6], Two-Stream GCN [7], Shift-GCN [8], MS-G3D Net [9] and Efficient GCN [10]. While these approaches achieve high accuracy on the training set, they often struggle to generalize to new scenarios and unseen categories. In addition, supervised methods require a large amount of labeled data, which are time-consuming and expensive to collect. They are also prone to overfitting, especially when the dataset is small or lacks diversity.

To address the above issues, self-supervised skeleton-based representation learning has gained popularity recently. Existing self-supervised approaches can be categorized into two paradigms: masked sequence modeling and contrastive learning. Masked sequence modeling leverages artificial supervision signals through pretext tasks like skeleton reconstruction [11], motion prediction [12], and skeleton colorization [13]. These methods utilize an encoder-decoder architecture to effectively capture the spatio-temporal dependencies within the skeleton sequence, enhancing the quality of representation learning. In contrast, contrastive learning-based methods [14], [15], often rely on negative samples, focus on learning discriminative instance-level representations between contrastive pairs. However, both types of approaches typically require additional components such as decoders or memory banks and involve intricate masking or sampling strategies. In addition, they typically focus on learning coarse-grained action representations, neglecting the fine-grained representations, which are crucial for dense prediction tasks.

Although dense prediction tasks in image or video understanding attract significant attention, they have not received adequate attention in skeleton-based action understanding. Compared to the action recognition task which is based on

- *H. Wang, W. Weng and X. Geng are with School of Computer Science and Engineering, Southeast University, Nanjing 211189, China, and also with Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China (email: hongsongwang@seu.edu.cn; 220232322@seu.edu.cn; xgeng@seu.edu.cn).*
- *J. Wang is with School of Software, Northwestern Polytechnical University, Xi'an 710072, China (email: jbwang@nwpu.edu.cn).*
- *F. Zhao is with State Key Laboratory for Novel Software Technology and School of Intelligence Science and Technology, Nanjing University, Nanjing 210023, China (email: zhaofang0627@gmail.com).*
- *G. Xie is with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China (gsxiehm@gmail.com).*
- *L. Wang is with New Laboratory of Pattern Recognition (NLPR), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), and also with School of Artificial Intelligence, University of Chinese Academy of Sciences (email: wangliang@nlpr.ia.ac.cn).*

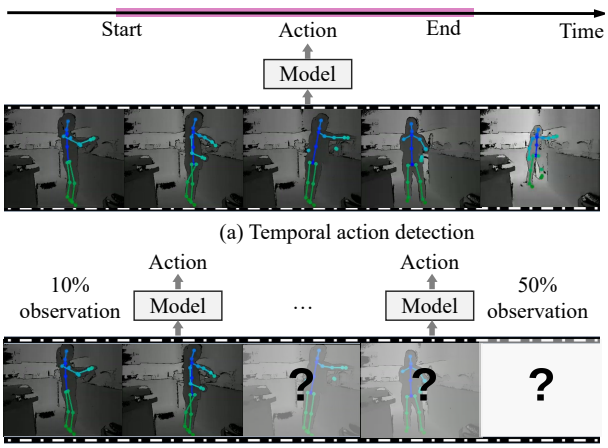


Fig. 1: Dense prediction tasks of skeleton-based action understanding. Temporal action detection directly processes raw long video sequences to identify both the action categories and their temporal boundaries, while action prediction requires continuously predicting the action in an online manner and recognizing it correctly as early as possible.

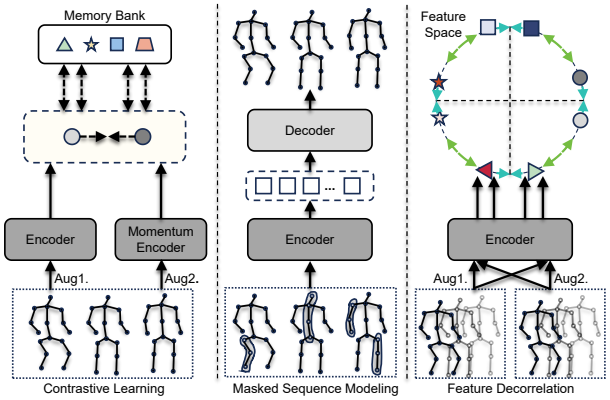


Fig. 2: Comparison of contrastive learning-based, masked sequence modeling-based, and feature decorrelation-based self-supervised skeleton-based representation learning paradigms. The objective of the feature decorrelation-based approach is to uniformly and consistently distribute samples within the representation space. Unlike masked sequence modeling, this method is lightweight, requiring neither a decoder nor complex masking strategies. Furthermore, it streamlines and simplifies contrastive learning by eliminating the need for a memory bank or an additional momentum encoder.

pre-segmented sequences, dense prediction tasks such as temporal action detection [5] and action prediction [16] are more aligned with real-world scenarios, as shown in Figure 1. A possible reason these problems have been overlooked is the lack of open-source benchmarks and well-recognized baselines.

Foundation models have recently made significant progress in image and video understanding [17]. For the area of skeleton-based action understanding, a foundation model should have the following characteristics: First, the model should be a simple, scalable, and easily trainable Transformer-based architecture. Second, it should be capable of learning both effective coarse-grained and fine-grained representations. Third, it should possess strong transferability across a variety of downstream tasks, including dense prediction tasks. However, most existing approaches for self-supervised skeleton-based representation learning do not meet the above requirements due to limited scalability and generalization. Dense prediction tasks, such as temporal action detection and action prediction, have been largely overlooked in existing works. Therefore, building a foundation model for an extensive set of tasks of skeleton-based human action understanding is an urgent problem that needs to be addressed.

To this end, we propose a foundation model with self-supervised dense representation learning named Unified Skeleton-based Dense Representation Learning (USDRL). Different from contrastive learning-based and masked sequence modeling-based methods, we use feature decorrelation for self-supervised action representation learning (see Figure 2). To learn effective dense representations, we design Multi-Grained Feature Decorrelation (MG-FD) which decorrelate features across temporal, spatial, and instance domains in a multi-grained manner, ensuring both consistency within individual samples and discriminability between different samples. Additionally, we propose a Transformer-based backbone, Dense Spatio-Temporal Encoder (DSTE), which is essential for capturing multi-grained features to generate dense representations, thus enhancing the model's capacity for effective dense prediction. The DSTE comprises two modules: Convolutional Attention (CA), which captures local feature relationships, and Dense Shift Attention (DSA), which uncovers hidden dependencies. Finally, we introduce a novel Multi-Perspective Consistency Training (MPCT) framework that incorporates explicit viewpoint cues and diverse skeleton data modalities to improve representation learning while maintaining efficiency. We categorize skeleton-based action understanding tasks into three main types: coarse prediction, dense prediction, and transferred prediction. Our approach is well-suited for dense prediction tasks, including action detection and prediction.

Our contributions can be summarized as follows:

- We present Unified Skeleton-based Dense Representation Learning (USDRL), a foundation model for skeleton-based action understanding that learns dense representations through multi-grained feature decorrelation.
- We design a novel Dense spatio-temporal encoder to model both the temporal and spatial streams, where each stream comprises Dense Shift Attention (DSA) and Convolutional Attention (CA). The DSA captures dense dependencies, while the CA excels at integrating local features.
- We conduct extensive experiments on 25 benchmarks across 9 downstream action understanding tasks, including recognition, retrieval, detection and prediction, demonstrating the effectiveness of our method.

The work is an extension of the published conference [18]. Compared to the preliminary version, the major differences are as follows: First, we propose a human skeleton-based

foundation model and adapt this model to different variants for dense prediction tasks, making it well suited for over eight action understanding tasks. Second, we introduce Multi-Perspective Consistency Training (MPCT), which includes both multi-view and multi-modal self-supervised training. Third, we conduct more comprehensive experiments and analysis, including additional three experiments on action prediction, action segmentation, and transfer learning for action retrieval.

2 Related Work

Self-Supervised Skeleton-Based Action Recognition: Existing works of self-supervised skeleton representation learning can be broadly categorized into two types: masked sequence modeling and contrastive learning. Masked sequence modeling methods [11], [12], [19] learn to represent skeletons through a pretext task that involves predicting the original sequence from masked and corrupted sequences. These tasks require predicting or reconstructing the original sequence from masked and corrupted sequences, thereby capturing the spatio-temporal dynamics of the actions.

Constrative learning-based methods can also be further subdivided into three categories: negative sample-based, self-distillation-based, and feature decorrelation-based. Negative sample-based methods [15], [20], [21] facilitate learning at the instance level by minimizing the differences with negative examples and maximizing those with positive examples within a memory bank. Self-distillation-based methods [22], [23] involves converting pairwise sample similarities into a probability distribution to emulate the teacher model's probability distribution, thus capturing crucial relational knowledge. Feature decorrelation [24] is a relatively novel approach. Compared to methods using negative samples and self-distillation, these approaches [18], [25] are more computationally efficient and cost-effective, requiring neither large batch size nor an additional momentum encoder and memory bank.

Skeleton-Based Action Detection/Localization: Temporal action detection or localization aims to identify the start time, end time, and category of each action instance in an untrimmed video. Common methods for skeleton-based action detection can be broadly categorized into two-stage and one-stage approaches. Most existing methods, such as [5], [26], [27], follow a two-stage pipeline. They first apply a frame-wise prediction approach and then generate a set of action proposals based on probabilistic matrices of action predictions for each frame. A few methods adopt a one-stage pipeline. For example, Sun et al. [28] propose a Transformer-based end-to-end baseline that integrates deformable attention mechanisms.

Skeleton-Based Action Segmentation: Temporal action segmentation predicts an action label to each frame of the video. Xu et al. [29] introduce a Connectionist Temporal Classification (CTC) loss to enhance the alignment of action segments and generate new labeled action sequences through motion interpolation. Li et al. [30] develop an improved spatial graph network to capture spatial dependencies and design a regression network to extract segmented encoding features and action boundary representations. Yang et al. [31] focus on learning skeleton representations in both video and frame spaces, improving the model's generalization for frame-wise action segmentation. Hyder et al. [32] use 2D skeleton heatmaps as input and apply Temporal Convolutional Networks (TCNs) to capture spatiotemporal dynamics. Ji et al. [33] incorporate language-assisted learning into skeleton-based action segmentation, utilizing linguistic knowledge to improve the understanding of relationships among joints and actions.

Skeleton-Based Action Prediction: Action prediction aims to recognize human actions from temporally incomplete video data. Most works focus on RGB-video based action prediction. For example, Aliakbarian et al. [34] introduce a novel loss function to encourage the network to predict the correct class as early as possible. Kong et al. [35] integrate LSTM with a memory module to record the discriminative information at early stage. Although skeleton-based action prediction is an important problem, it has not received sufficient attention. Liu et al. [16] propose a network architecture based on 1D dilated convolutions to flexibly handle variations in the temporal scale of observed actions. Wang et al. [36] propose a teacher-student network and a cross-task progressive distillation knowledge training method to improve early-stage action prediction performance.

Unified Representation Learning from Skeletons: Unified pretraining methods in skeleton-based representation learning first train models on specific pretext tasks before adapting them to diverse downstream applications, improving their flexibility and effectiveness. MotionBERT [19] utilizes masked sequence modeling with a 2D-to-3D lifting strategy for pretraining, followed by fine-tuning on multiple tasks. SkeletonMAE [11] employs a similar masked modeling approach but focuses on direct sequence reconstruction. PCM^3 [27] integrates both contrastive learning and masked modeling, leveraging their combined benefits to enhance generalization. Skeleton-in-Context [37] introduces an in-context skeleton sequence modeling framework, capturing spatial relationships and contextual dependencies to refine representations. UmURL [38] formulates its pretraining objective around feature decorrelation, effectively incorporating multiple skeleton modalities. Unified Pose Sequence (UPS) [39] unifies heterogeneous output formats by representing both text-based action labels and coordinate-based human poses as language sequences. However, UPS targets four pose-based tasks: 3D action recognition, 2D action recognition, 3D pose estimation, and 3D early acton prediction, while our approach addresses an extensive set of action understanding tasks. Large Language Model as an Action Recognizer (LLM-AR) [40] is also proposed; however, LLM-AR focuses solely on the action recognition task. Wang et al. [41] investigate data heterogeneity in human skeletons and propose a unified framework for learning action representation from different heterogeneous skeletons. Although these techniques perform well in sequence-level tasks, they struggle with frame-level dense prediction. To bridge this gap, we propose a novel skeleton-based dense representation learning framework centered on feature decorrelation, aiming to improve fine-grained feature capture.

Feature Decorrelation: Feature decorrelation is a technique in self-supervised learning designed to prevent model collapse by minimizing redundancy across feature dimensions and ensuring diverse representations. Traditionally, contrastive

learning with negative samples has been used to tackle this issue, but it often requires additional components such as a momentum encoder, a large memory bank, and substantial batch sizes, making it resource-intensive. To provide an alternative, W-MSE [42] introduces a decorrelation-based approach that does not rely on negative samples. In W-MSE, positive sample pairs (*e.g.*, different augmentations of the same image) are processed through a shared encoder, followed by a whitening module that standardizes feature distributions and ensures linear independence across dimensions. Barlow Twins [24] employs a related strategy but optimizes the cross-correlation matrix between two encoded representations, encouraging it to approximate an identity matrix to achieve the same goal. Building on this idea, VICREG [25] stabilizes self-supervised learning by incorporating variance, invariance, and covariance to prevent collapse. In this work, we integrate Barlow Twins and VICREG principles to develop a self-supervised skeleton-based representation learning framework using feature decorrelation.

Video Foundation Models: Video Foundation Models aim to develop a general-purpose representation for diverse video understanding tasks. Our review focuses solely on works that achieve effective representation learning through pre-training with specific loss functions, broadly classified into discriminative and generative approaches. Typical discriminative pretraining are Video-Text Contrastive [43] and Video-Text Matching [44]. The former pulls similar representations together while pushing dissimilar ones apart, whereas the latter aims to maximize the matching score for a given video-text pair. Generative approaches aim to reconstruct masked information within video data. Notable examples include Masked Language Modeling [45], Mask Video Modeling [46], and Mask Image Modeling [47].

Foundational Motion Models: There are some foundational motion models that unify human motion prediction and synthesis, trajectory prediction, 3D pose estimation, and action understanding, e.g., Unified Pose Sequence (UPS) [39], SkeletonMAE [11], MotionBERT [19], Sports-Traj [48] and Multi-modal Large Motion Model [49]. While motion synthesis emphasizes generating new motions consistent with semantic input, motion understanding focuses on interpreting and analyzing motions.

3 METHODS

We introduce a foundation model for skeleton-based human action understanding. As depicted in Figure 3, this model follows the standard two-stream pipeline [4] and primarily consists of a Dense Spatio-Temporal Encoder (DSTE). During training, three specialized projectors, namely temporal, spatial, and instance projectors, each composed of multiple linear layers, are employed. Additionally, a self-supervised training paradigm called Multi-Grained Feature Decorrelation (MG-FD) is proposed. Unlike contrastive methods relying on negative samples [15], this technique eliminates the need for a momentum encoder or memory bank.

During the forward pass, an augmented 3D skeleton sequence $\mathbf{X} \in \mathbb{R}^{C_{in} \times T \times V \times M}$ of length T with V skeletal vertices is first transformed into two distinct views: temporal and spatial. The sequence is reshaped into $\mathbf{X}_t \in \mathbb{R}^{T \times (M \times V \times C_{in})}$ for the temporal domain and $\mathbf{X}_s \in \mathbb{R}^{(M \times V) \times (T \times C_{in})}$ for the spatial domain. These rearranged sequences are then mapped into an embedding space, producing feature representations $\mathbf{F}_t \in \mathbb{R}^{T \times C_e}$ and $\mathbf{F}_s \in \mathbb{R}^{V \times C_e}$. Next, these embeddings pass through the DSTE's respective temporal and spatial streams, generating dense representations $\mathbf{y}_t \in \mathbb{R}^{T \times C_r}$ and $\mathbf{y}_s \in \mathbb{R}^{V \times C_r}$. Applying MaxPooling condenses these into vectors $\mathbf{y}_t, \mathbf{y}_s \in \mathbb{R}^{C_r}$. These are subsequently projected into a new space using the three domain-specific projectors, yielding projection vectors $\mathbf{z}_t, \mathbf{z}_s \in \mathbb{R}^{C_p}$ and $\mathbf{z}_i \in \mathbb{R}^{2 \times C_p}$. Here, C_{in}, C_e, C_r, C_p represent the channel dimensions for input, embedding, representation, and projection, respectively. Finally, the proposed MG-FD, $\mathcal{L}_{mfd}$, is computed among the projection vectors to enforce decorrelation constraints across temporal, spatial, and instance features.

3.1 Dense Spatio-Temporal Encoder

The proposed Dense Spatio-Temporal Encoder (DSTE) consists of two parallel temporal and spatial streams designed to capture dynamic and structural features, respectively. As illustrated in Figure 4, both streams are constructed using multiple stacked layers, each integrating a Dense Shift Attention (DSA) module and a Convolutional Attention (CA) module. The architecture of a single DSTE layer is detailed below.

Dense Shift Attention (DSA): Given an input embedding sequence $\mathbf{F} \in \mathbb{R}^{L \times C_e}$ in either the temporal or spatial domain, the DSA module leverages an MLP with two learnable weight matrices, $W_1, W_2 \in \mathbb{R}^{L \times L}$, to uncover underlying relationships between embeddings. Here, L represents the sequence length. Specifically, the sequence is first reshaped into $\mathbf{F}_1 \in \mathbb{R}^{C_e \times L}$, then processed as follows:

$$\mathbf{F}_h = \mathrm{ReLU}\left(W_1 \mathbf{F}_1\right) W_2 + \mathbf{F}_1. \tag{1}$$

The transformed representation $\mathbf{F}_h$, enriched with global context, is then combined with the original sequence $\mathbf{F}$ through a DenseShift operation. This mechanism enables embeddings to incorporate semantic information from the entire sequence:

$$\mathbf{F}_m = \mathbf{Mask} \odot \mathbf{F}_h + \overline{\mathbf{Mask}} \odot \mathbf{F}, \tag{2}$$

where $\mathbf{Mask}$ is a binary vector with every gap-th element set to 1 and all others set to 0, while $\overline{\mathbf{Mask}} = 1 - \mathbf{Mask}$.

Subsequently, $\mathbf{F}_m$ and $\mathbf{F}$ undergo sparse Self-Attention (SA) and Feed-Forward Network (FFN) transformations. The final output of the DSA module, $\mathbf{F}_d$, is obtained via:

$$\mathbf{F}_d = \mathrm{FFN}\left(\mathrm{SA}\left(\mathbf{F}_m\right)\right) + \mathrm{FFN}\left(\mathrm{SA}\left(\mathbf{F}\right)\right), \tag{3}$$

where $\mathbf{F}_d$ is the output of the DSA module.

Convolutional Attention (CA): The CA module begins by applying 1D channel-wise temporal or spatial convolutions on the input embeddings $\mathbf{F} \in \mathbb{R}^{T/V \times C_e}$ to enhance local feature interactions. The resulting representations then pass through a self-attention mechanism, capturing long-range dependencies and global patterns. The final output $\mathbf{F}_g \in \mathbb{R}^{T/V \times C_r}$ is computed as follows:

$$\mathbf{F}_g = \mathrm{FFN}\left(\mathrm{SA}\left(\mathrm{Conv}\left(\mathbf{F}\right) + \mathbf{F}\right)\right), \tag{4}$$

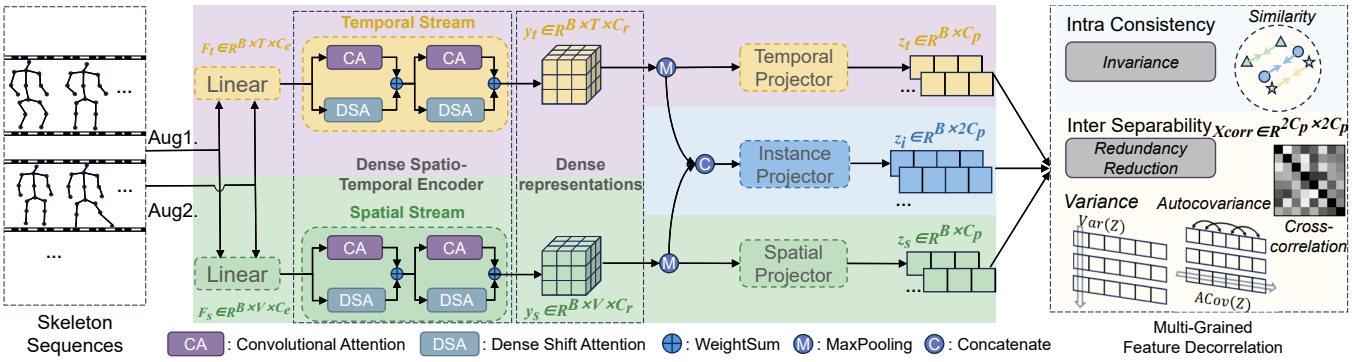


Fig. 3: The proposed Unified Skeleton-based Dense Representation Learning (USDRL) framework. USDRL incorporates a two-stream architecture with the Dense Spatio-Temporal Encoder (DSTE). The DSTE processes skeleton sequences to derive dense representations, which are further refined through MaxPooling and concatenation to generate condensed vectors. The Multi-Grained Feature Decorrelation is devised to mitigate model collapse and guarantee both intra-sample consistency and inter-sample separability. To further enhance robustness across different viewpoints and facilitate multimodal learning, a multi-perspective consistency training strategy is used during training.

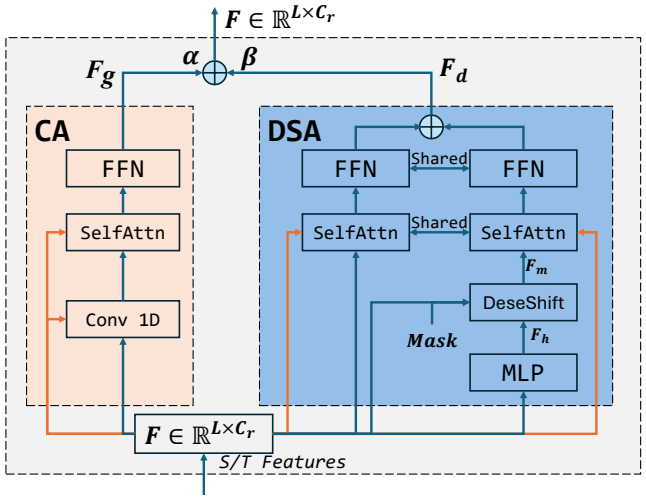


Fig. 4: The fundamental layer of the Dense Spatio-Temporal Encoder. It consists of the Convolutional Attention (CA) and Dense Shift Attention (DSA) modules, where $\oplus$ represents a weighted sum operation.

where FFN, SA denote Feed-Forward layer and Self-Attention layer, respectively.

Spatio-Temporal Representations: The outputs of the CA and DSA modules are combined using a weighted sum:

$$\mathbf{y} = \alpha\,\mathrm{CA}\left(\mathbf{F}\right) + \beta\,\mathrm{DSA}\left(\mathbf{F}\right), \tag{5}$$

where α and β are weighting coefficients satisfying $\alpha + \beta = 1$. The functions $\mathrm{DSA}(\cdot)$ and $\mathrm{CA}(\cdot)$ refer to the operations defined in Eq. 3 and Eq. 4, respectively.

The final dense spatio-temporal representations, $\mathbf{y}_t \in \mathbb{R}^{T\times C_r}$ and $\mathbf{y}_s \in \mathbb{R}^{V\times C_r}$, serve as the core feature representations:

$$\mathbf{y}_s, \mathbf{y}_t = \mathrm{DSTE}\left(\mathbf{F_s}, \mathbf{F_t}\right), \tag{6}$$

where $\mathrm{DSTE}(\cdot)$ denotes the Dense Spatio-Temporal Encoder. These representations effectively preserve both the temporal dynamics and spatial structure of the skeleton sequence.

3.2 Multi-Grained Feature Decorrelation

After mapping the learned action representations to a higher-dimensional space via three specialized projectors, a Multi-Grained Feature Decorrelation training loss is formulated, specifically designed for the temporal, spatial, and instance domains. The overall training loss $\mathcal{L}$ is expressed as:

$$\mathcal{L} = \mathcal{L}_{fd}\left(\mathbf{Z}\right) + \tau\left(\mathcal{L}_{fd}\left(\mathbf{Z}_s\right) + \mathcal{L}_{fd}\left(\mathbf{Z}_t\right)\right), \tag{7}$$

where $\mathbf{Z_s}$, $\mathbf{Z_t}$, and $\mathbf{Z}$ represent feature matrices containing high-dimensional representations for spatial, temporal, and instance domains, respectively. Here, $\mathcal{L}_{fd}$ signifies the feature decorrelation loss, and τ is a hyperparameter that regulates the contributions of spatial and temporal components.

The feature decorrelation loss incorporates both Intra-Sample Consistency and Inter-Sample Separability. We elaborate on these components in the context of the instance domain as an example.

Intra-Sample Consistency: Since data augmentation preserves intrinsic information, projection vectors $\mathbf{z}_k$ derived from K augmentations should retain semantic consistency. To enforce this, we introduce $\mathcal{L}_{con}$, comprising a *Similarity term* and an *Invariance term*. The Similarity term minimizes the mean squared error (MSE) among augmentations, ensuring close proximity of representations, whereas the Invariance term aligns their autocorrelation towards 1, promoting consistency. $\mathcal{L}_{con}$ is defined as:

$$\mathcal{L}_{con} = \frac{1}{K}\sum_{a=1}^{K}\Big(\kappa\underbrace{\left\|\mathbf{z}_a - \overline{\mathbf{z}}\right\|_2}_{\text{Similarity}} + \eta\sum_{b=1|b\neq a}^{K}\underbrace{\mathrm{tr}\left(\mathbf{I} - \widehat{\mathbf{z}}_a^{\mathrm{T}}\widehat{\mathbf{z}}_b\right)}_{\text{Invariance}}\Big), \tag{8}$$

where $\widehat{\mathbf{z}}$ is the normalized vector, $\overline{\mathbf{z}}$ is the average across augmentations, tr is the trace operator, $\mathbf{I}$ denotes the identity matrix, and η, κ are weighting factors.

Inter-Sample Separability: Without inter-sample separability, representations may become highly correlated, leading to redundancy and model collapse. Inspired by recent self-supervised learning advancements, we define the Inter-Sample Separability loss $\mathcal{L}_{sep}$ using three complementary terms: variance, covariance, and cross-correlation, which

together ensure feature decorrelation and distinguishability across samples.

Variance Term. Given a matrix $\mathbf{Z} \in \mathbb{R}^{N \times C_p}$, where N is the batch size and C_p represents feature dimensions, the variance term maintains feature diversity by ensuring each dimension has a variance exceeding a threshold γ. A small constant ϵ prevents instability:

$$V(\mathbf{Z}) = \frac{1}{C_p} \sum_{j=1}^{C_p} \text{ReLU}\Big(\gamma - \sqrt{Var(\mathbf{Z}_{:,j}) + \epsilon}\Big), \tag{9}$$

where $Var(\mathbf{Z}_{:,j})$ denotes variance along dimension j.

Auto-Covariance Term. To promote independence among features, the auto-covariance term minimizes cross-dimension correlations within $\mathbf{Z}$:

$$AC(\mathbf{Z}) = \frac{1}{C_p} \sum_{i=1}^{C_p} \sum_{j=1|j\neq i}^{C_p} [ACov(\mathbf{Z})]_{i,j}^2, \tag{10}$$

where $ACov(\mathbf{Z})$ is the auto-covariance matrix.

Cross-Correlation (Xcorr) Term. This term reduces redundancy by decorrelating features among different augmentation versions. Given two augmented feature sets $\mathbf{Z}_a$ and $\mathbf{Z}_b$, we compute their cross-correlation matrix. The off-diagonal elements should ideally be close to zero, ensuring minimal redundancy:

$$XC(\mathbf{Z}_a, \mathbf{Z}_b) = \sum_{i=1}^{C_p} \sum_{j=1|j\neq i}^{C_p} [Xcorr(\mathbf{Z}_a, \mathbf{Z}_b)]_{i,j}^2, \tag{11}$$

where $Xcorr(\mathbf{Z}_a, \mathbf{Z}_b)$ is the cross-correlation matrix.

By integrating these terms, the final Inter-Sample Separability loss $\mathcal{L}_{sep}$ is formulated as:

$$\mathcal{L}_{sep} = \sum_{a=1}^{K} \Big(\mu V(\mathbf{Z}_a) + AC(\mathbf{Z}_a) + \lambda \sum_{b=a+1}^{K} XC(\mathbf{Z}_a, \mathbf{Z}_b)\Big), \tag{12}$$

where μ and λ balance the contributions of different terms. Finally, the instance domain loss is given by:

$$\mathcal{L}_{fd}(\mathbf{Z}) = \mathcal{L}_{con}(\mathbf{Z}) + \mathcal{L}_{sep}(\mathbf{Z}). \tag{13}$$

3.3 Multi-Perspective Consistency Training

Multi-View Training: For self-supervised training, most works treat different data-augmented versions of the same sample as positive samples to learn augmentation-invariant features. As acquiring multiple viewpoints of the same action sequence is relatively easy, we explicitly incorporate viewpoint information for self-supervised action representation learning. Positive pairs are constructed not only via data augmentation, but also from the same action sequence captured from different viewpoints. This strategy encourages the model to learn representations that are both augmentation-invariant and viewpoint-invariant. Specifically, inputs for the unified and single-modality branches are derived from the same action sequence recorded by different cameras, with each view subjected to distinct data augmentations. Although the single-modality inputs originate from the same underlying sequence, they undergo separate augmentations to enhance input diversity and reduce reliance on low-level cues during inter-modal alignment. Through this approach, the model is guided to focus on high-level semantic features of the action itself, while mitigating the influence of low-level discrepancies caused by viewpoint variations.

Multi-Modal Training: Different modalities [5], such as joints, bones, and motion, can be used to represent a given skeleton sequence. Most existing works combine multimodal skeleton results using late fusion, which employs three separate streams and increases the computational cost by three times. Let $\mathbf{X^k}$ represent a specific modality, where k is the index of the various modalities, i.e., $k \in \{joint, bone, motion\}$. The input from different modalities is mapped to a common high-dimensional embedding space through distinct linear embedding modules to obtain $\mathbf{F^k}$. To reduce computational complexity, early fusion is employed to integrate embeddings from different modalities into a unified representation. For both the spatial and temporal streams, the multimodal dense representation is computed as:

$$\mathbf{y} = \text{DSTE}(\text{Fusion}(\mathbf{F}^{joint}, \mathbf{F}^{bone}, \mathbf{F}^{motion})), \tag{14}$$

where $\mathbf{y}$ denotes the fused dense representation of joint, bone, and motion modalities, and $\text{Fusion}(\cdot)$ denotes the early fusion operations, such as averaging.

For the purposes of simplicity and efficiency, our approach omits the inter-modal consistency loss when compared with UmURL [38]. As multimodal features share the common encoder backbone, early fusion significantly reduces model parameters in comparison to other approaches that maintain separate backbones for multimodal inputs.

3.4 Adaptation for Dense Prediction Tasks

Action Detection: For temporal action detection, we follow the two-stage pipeline [5] that considers action detection as a frame-wise classification problem, with an additional action class representing the background. During inference, for long videos, we adopt a sliding window approach to sample fixed-length segments, feed them into the network, and then concatenate the predictions along the temporal dimension. A post-processing step is used to transform the predicted frame-wise probabilities into a set of triplets, each consisting of a start frame, an end frame, and an action category.

Action Segmentation: Skeleton-based action segmentation aims to predict the action category for each frame. We follow the end-to-end pipeline [50] that directly feeds the long sequence into the network.

Action Prediction: Action prediction aims to recognize the action as early as possible given a partially observed action sequence. For such task that requires temporal causality, we employ causal attention operations in the transformer for the temporal stream.

4 Experiment

4.1 Datasets

NTU-RGB+D 60 Dataset (NTU-60) [51] is a dataset for skeleton-based human action recognition, featuring 56,880 sequences, where each sequence represents a single action performed by one of 40 subjects across 60 categories. Each action is captured as 3D coordinates of 25 body joints. We employ the established Cross-Subject (X-sub) and Cross-View (X-view) evaluation protocols.

NTU-RGB+D 120 Dataset (NTU-120) [52] is an extended version of NTU-60, featuring 114,480 sequences across 120 action categories performed by 106 subjects. Each action, represented by 3D coordinates of 25 body joints, is evaluated using Cross-Subject (X-sub) and Cross-Setup (X-set) protocols. In the X-sub, half of the subjects are used for training and the rest for testing. The X-set splits the data from 32 camera setups for diverse training and testing.

PKU Multi-Modality Dataset (PKUMMD) [53] is a large-scale, multi-modality dataset for 3D human action understanding, with two subsets: I and II. PKUMMD I is dedicated to action detection, using continuous video streams to identify and recognize specific actions. PKUMMD II focuses on tasks of action recognition and segmentation. The recognition crops action instances based on temporal annotations and aims to classify them into 51 distinct categories using a cross-subject protocol, while the segmentation aims to directly segment the raw skeleton sequence. This dataset presents significant challenges due to variations in camera angles and subject positioning, comprising 5,332 training samples and 1,613 testing samples.

UAV-Human [54] is a large-scale human action understanding dataset captured by unmanned aerial vehicles. We use the 2D skeleton sequences for action recognition.

4.2 Implementation Details

Model Structure: For data process and augmentation, we adhere to the same strategies employed in works [38], [67]. The DSTE Encoder is a two-layer architecture. The channel dimensions of embedding, representation, and projection are configured as 1024, 1024, and 2048, respectively. The projector consists of two linear layers, each followed by batch normalization and a ReLU activation, and a third linear layer. The output dimensions for the spatial, temporal, and instance projectors are 2048, 2048, and 4096, respectively.

Our model comprises a skeleton encoder serving as the backbone, accompanied by three domain-specific projectors, each with a simple and identical structure consisting of several linear layers. We employ both STTR and DSTE as alternative backbones for our model, facilitating comparative analysis, each of these backbones consists of a two-layer architecture. For the NTU60/120 and PKU-MMD I datasets, the channel numbers of embedding, representation, and projection, C_e, C_r, and C_p, are set to 1024, 1024, and 2048, respectively. In contrast, for the PKU-MMD II dataset, which has fewer samples yet features longer, more complex action sequences than the NTU datasets, the dimensions are adjusted to 512, 512, and 1024 respectively.

Hyperparameters in Pre-training: During the pretraining stage, the gap in the Dense Shift operation is set to 4. The hyperparameters for the weighted sum of the CA and DSA modules, α and β, are both set to 0.5. the hyperparameters τ, used to balance the different domain terms in the pretrain loss $\mathcal{L}_{mfd}$, is also set to 0.5. In Intra-sample Consistency, the hyperparameters κ and η are set to 5 and 5e-4 respectively. In the case of Inter-sample separability, the hyperparameters μ and λ, used to balance the cross-correlation and variance terms, are set to 1 and 0.001 respectively.

Training Details: For the optimizer, we employ the Adam with a weight decay of 1e-5. The batch size is set to 324. In practice, we observe that batch sizes of 256 and 512 yield similar performance metrics. However, larger batch sizes significantly reduce training time, suggesting a trade-off between computational efficiency and resource utilization. The model undergoes training for 450 epochs on the NTU60/120 datasets and 1200 epochs on the PKU-MMD II dataset. The initial learning rate is set to 5e-4 and reduced to 5e-5 at epoch 350 for NTU60/120 and epoch 1000 for PKU-MMD II, respectively.

Baselines and Evaluations: We consider the transformer-based STTR [76] with self-supervised training and plain feature decorrelation as the baseline. We gradually replace the training method with the proposed MG-FD and the backbone with the proposed DSTE.

We evaluate the effectiveness of our method across eight downstream tasks, which are categorized into coarse prediction, dense prediction, and transfer prediction. Coarse prediction tasks include both unsupervised and semi-supervised action recognition, as well as action retrieval. The instance-level representation is obtained by applying a MaxPooling operation across frames and concatenating the spatial and temporal streams. Dense prediction tasks include action detection, segmentation and prediction, where frame-wise predictions are made using dense representations. Transferred prediction tasks involve transfer learning for both action recognition and retrieval. Since transfer learning for skeleton-based action retrieval has not been studied before, we set up a benchmark and implement several baselines to achieve this goal.

4.3 Results of Coarse Prediction Tasks

Coarse prediction uses the learned global representation for action recognition and retrieval. We conduct experiments on three popular tasks: unsupervised action recognition, semi-supervised action recognition and skeleton-based action retrieval.

Unsupervised Action Recognition: This task involves training an encoder for skeleton-based action recognition through unsupervised pre-training. After the pre-training phase, a fully connected layer is appended to the encoder's backbone network. During the evaluation process, the backbone network remains frozen, and only this newly added linear layer is fine-tuned to assess the quality of the representations learned during pre-training. Table 1 presents accuracies on the popular five benchmarks from NTU-60, NTU-120, and PKU-MMD II datasets. The comparative methods are classified into four groups according to different self-supervised learning methods: hybrid learning, masked sequence modeling, constrastive learning and feature decorrelation. The proposed USDRL achieves state-of-the-art performance across all types of comparative approaches, surpassing the previous state-of-the-art feature decorrelation-based method by an average margin of 2.8% across all benchmarks. Specifically, USDRL utilizing only the joint modal consistently outperforms UmURL [38] employing three modalities by approximately 1.4%. In addition, the improved USDRL with multi-perspective consistency training outperforms the preliminary method by 1% on the NUT-60 dataset and 2.6% on the PKU-MMD II dataset. It can be concluded that, when compared with other self-supervised

TABLE 1: Comparison of unsupervised action recognition results. J: Joint, M: Motion, B: Bone.

Method	Publisher	Modality	NTU-60		NTU-120		PKU-MMD II
			x-sub	x-view	x-sub	x-set	x-sub
Masked Sequence Modeling							
MS^{2}L [55]	ACM MM'20	J	52.6	-	-	-	27.6
3s-Colorization [56]	ICCV'21	J	75.2	83.1	-	-	-
GL-Transformer [57]	ECCV'22	J	76.3	83.8	66	68.7	-
Masked Colorization [13]	TPAMI'23	J	79.1	87.2	69.2	70.8	49.8
PCM^3 [27]	ACM MM'23	J	83.9	90.4	76.5	77.5	51.5
SkeletonMAE [58]	ICMEW'23	J	74.8	77.7	72.5	73.5	36.1
MAMP [12]	ICCV'23	J	84.9	89.1	78.6	79.1	53.8
MacDiff [59]	ECCV'24	J	86.4	91.0	79.4	80.2	-
MMFR [60]	TCSVT'24	J	84.2	89.5	77.1	78.8	54.4
Contrastive Learning							
AimCLR [15]	AAAI'22	J	74.3	79.7	63.4	63.4	38,5
CMD [22]	ECCV'22	J	79.8	86.9	70.3	71.5	43.0
CPM [14]	ECCV'22	J	78.7	84.9	68.7	69.6	48.3
PSTL [61]	AAAI'23	J	77.3	81.8	66.2	67.7	49.3
HaLP [62]	CVPR'23	J	79.7	86.8	71.1	72.2	43.5
HiCo [63]	AAAI'23	J	81.1	88.6	72.8	74.1	49.4
2s-DMMG [64]	TIP'23	J+M	84.2	89.3	72.7	72.4	58.8
Skeleton-logoCLR [65]	TCSVT'24	J	82.4	87.2	72.8	73.5	54.7
KTCL [66]	TMM'24	J	82.4	89.4	74.4	74.5	55.5
SCD-Net [67]	AAAI'24	J	86.6	91.7	76.9	80.1	54.0
ViA [68]	IJCV'24	J+M	78.1	85.8	69.2	66.9	-
IGM [69]	ECCV'24	J	86.2	91.2	80.0	81.4	-
Feature Decorrelation							
HYSP [70]	ICLR'23	J	78.2	82.6	61.8	64.6	-
UmURL [38]	ACM MM'23	J	82.3	89.8	73.5	74.3	52.1
UmURL [38]	ACM MM'23	J+M+B	84.2	90.9	75.2	76.3	52.6
Heter-Skeleton [41]	CVPR'25	J	80.2	88.0	70.7	73.5	47.7
USDRL (STTR)	Preliminary work	J	84.2	90.8	76.0	76.9	51.8
USDRL (DSTE)	Preliminary work	J	85.2	91.7	76.6	78.1	54.4
USDRL (STTR)	This work	J+M+B	85.8	91.8	77.5	78.8	54.7
3s-ensemble							
3s-CMD [22]	ECCV'22	J+M+B	84.1	90.9	74.7	76.1	52.6
3s-CPM [14]	ECCV'22	J+M+B	83.2	87.0	73.0	74.0	51.5
3s-HiCLR [23]	AAAI'23	J+M+B	80.4	85.5	70.0	70.4	-
3s-SkeAttnCLR [71]	IJCAI'23	J+M+B	82.0	86.5	77.1	80.0	55.5
3s-SSRL [72]	TCSVT'23	J+M+B	81.6	85.1	69.2	71.5	50.2
3s-PCM^3 [27]	ACM MM'23	J+M+B	87.4	93.1	80.0	81.2	58.2
3s-ActCLR [73]	CVPR'23	J+M+B	84.3	88.8	74.3	75.7	-
3s-RVTCLR+ [74]	ICCV'23	J+M+B	79.7	84.6	68.0	68.9	-
3s-PSTL [61]	AAAI'23	J+M+B	79.1	82.6	69.2	70.3	52.3
3s-CSTCN [75]	TMM'23	J+M+B	85.8	92.0	77.5	78.5	53.9
3s-UmURL [38]	ACM MM'23	J+M+B	84.4	91.4	75.8	77.2	54.3
3s-USDRL (DSTE)	This work	J+M+B	**87.1**	**93.2**	**79.3**	**80.6**	**59.7**

learning methods, such as *Mask Sequence Modeling*, *Negative-based Contrastive Learning*, and *Hybrid Learning*, USDRL exhibits superior performance. Furthermore, training USDRL with additional bone and motion streams and subsequently conducting an ensemble of these models leads to a further substantial improvement.

Semi-Supervised Action Recognition: In the semi-supervised setting, the pre-trained encoder is first loaded, and subsequently, the entire model is fine-tuned using only 1% and 10% of randomly sampled labeled training data. Results on the NTU-60 dataset are reported in Table 2.

In the scenario where a mere 1% of the training data is labeled, our proposed method attains accuracies of 57.3% and 60.7% on the x-sub and x-view evaluation protocols, respectively. Specifically, when leveraging 1% of labeled training samples, our method surpasses the current state-of-the-art approaches by a margin of 4.3% on the x-view protocol. Moreover, when utilizing 10% of labeled data for training, it further extends this performance advantage, outperforming the existing best methods by 3.7% on the same x-view protocol. These results confirm the robust generalization capability of our approach and demonstrate its competitive performance in semi-supervised action recognition.

TABLE 2: Comparison of performance under semi-supervised evaluation protocol on the NTU60 dataset.

Method	x-sub		x-view	
	1 % data	10 % data	1 % data	10 % data
MS^{2}L [55]	33.1	65.2	-	-
ISC [77]	35.7	65.9	38.1	72.5
HiCLR [23]	51.1	74.6	50.9	79.6
CMD [22]	50.6	75.4	53	80.2
PCM^3 [27]	53.8	77.1	53.1	82.8
HiCo [63]	54.4	73.0	54.8	78.3
Heter-Skeleton [41]	55.0	76.3	55.0	79.1
USDRL (STTR)	55.0	76.1	59.1	82.0
USDRL (DSTE)	**57.3**	**80.2**	**60.7**	**84.0**

2D Skeleton-Based Action Recognition: In addition to 3D

TABLE 3: Results of 2D skeleton-based action recognition on the UAV-Human dataset. S and U denote supervised and unsupervised training, respectively.

Method	Modality	Training	CS-v1	CS-v2
ST-GCN [6]	J	S	30.2	56.1
2s-AGCN [7]	J+B	S	34.8	66.7
HARD-Net [78]	J	S	37.0	-
Shift-GCN [8]	J	S	38.0	67.0
LLM-AR [40]	J	S	46.3	-
USDRL (STTR)	J	U	31.7	50.2
USDRL (DSTE)	J	U	36.3	60.8

TABLE 4: Comparison of action retrieval results. * denotes the improved method of our USDRL.

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-set
ISC [77]	62.5	82.6	50.6	52.3
HaLP [62]	65.8	83.6	55.8	59
CMD [22]	70.6	85.4	58.3	60.9
UmURL [38]	71.3	88.3	58.5	60.9
PCM^3 [27]	73.7	88.8	63.1	**66.8**
HiCo [63]	68.3	84.8	56.6	59.1
Heter-Skeleton [41]	66.3	87.1	55.7	59.8
USDRL (STTR)	73.7	88.5	58.9	64.8
USDRL (DSTE)	75.0	89.3	63.3	66.7
USDRL (STTR)*	74.4	93.5	62.4	65.6

skeleton-based action recognition, we evaluate our approach for 2D skeleton-based action recognition. Results on the challenging UAV-Human [54] is shown in Table 3. It can be observed that our method even outperforms some supervised training methods.

Skeleton-Based Action Retrieval: In this task, representations extracted from the pre-trained encoder are directly employed for retrieval tasks, obviating the need for any additional training. Specifically, the nearest neighbor is identified within the representation space by calculating the cosine similarity between the query representation and all the gallery representations. Results on the NTU-60 and NTU-120 datasets are presented in Table 4, where our method is compared with various state-of-the-art approaches. By only leveraging the joint modality, our proposed model substantially outperforms previous works [27], [38], thereby underscoring the discernibility and efficacy of the features acquired through our approach.

TABLE 5: Comparsion of action detection results on PKU-MMD I xsub benchmark with an overlap ratio of 0.5.

Method	mAP$_a$ (%)	mAP$_v$ (%)
MS^{2}L [55]	50.9	49.1
CRRL [79]	52.8	50.5
ISC [77]	55.1	54.2
CMD [22]	59.4	59.2
PCM^3 [27]	61.8	61.3
USDRL (STTR)	66.1	65.9
USDRL (DSTE)	**75.7**	**74.9**

TABLE 6: Comparsion of action prediction results on the NTU-60 dataset.

Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DeepSCN [80]	16.8	21.5	30.5	39.9	48.7	54.6	58.2	60.2	60.0	58.6
MSRNN [81]	15.2	20.3	29.5	41.4	51.6	59.2	63.9	67.4	68.9	69.2
P-TSL [36]	27.8	35.8	46.3	58.5	67.4	73.9	77.6	80.1	81.5	82.0
USDRL (STTR)	24.7	36.5	54	65.7	72.8	76.9	80.3	82.4	83.7	84.2
USDRL (DSTE)	25.5	36.8	54.8	66.7	73.6	77.9	81.4	83.6	84.5	85.2

TABLE 7: Comparison of action segmentation results on the PKU-MMD II dataset.

Method	Acc	Edit	F10	F25	F50
ST-GCN [6]	64.9	-	-	-	15.5
MS-TCN [82]	65.5	-	-	-	46.3
ETSN [83]	68.4	67.1	70.4	65.5	52.0
CTC [29]	69.2	-	69.9	66.4	53.8
DeST [50]	67.6	66.3	71.7	68	55.5
USDRL (STTR)	68.7	67.5	70.9	67.3	56.2

4.4 Results of Dense Prediction Tasks

Dense prediction aims to learn frame-wise representation for skeleton-based action understanding. We perform experiments on three important tasks: action detection, action segmentation and action prediction.

Skeleton-Based Action Detection: We evaluate the performance of our model for action detection on the PKU-MMD I dataset. Following the standard protocols delineated in [26], [53], we adopt two metrics: the mean average precision (mAP) for different actions (denoted as mAP$_a$) and the mAP for different videos (denoted as mAP$_v$). These metrics are computed with a default overlap ratio threshold of 0.5. As presented in Table 5, our approach achieves a remarkable performance improvement, significantly surpassing the state-of-the-art methods on this dataset. For example, when employing the same backbone of STTR, our approach outperforms CMD [22] and PCM^3 [27] by 6.7% and 4.3% in terms of mAP$_a$, respectively. Furthermore, our approach with the proposed DSTE backbone also surpasses the variant using STTR by nearly 9%. These results accentuate the model's adeptness in acquiring dense and discriminative feature representations, which are pivotal for precise action detection.

Skeleton-Based Action Prediction: For action prediction, we report the action recognition accuracies with the observation ratio ranging from 10% to 100% in Table 6. The state-of-the-art P-TSL [36] trains 10 distinct offline action recognition models, each corresponding to one of 10 different observation ratios. In contrast, our causal model employs a single unified architecture for online frame-wise action recognition, offering superior scalability and efficiency compared to existing offline models. The predicted probabilities of observed frames are simply aggregated for action recognition. Although our approach initially underperforms the offline state-of-the-art P-TSL [36] at an observation ratio of 10%, it surpasses P-TSL's performance when the observation ratio reaches 20% or higher. Specifically, it outperforms P-TSL [36] by 6.2% absolute accuracy when the observation ratio is 50%, which demonstrates the potential of our model for online action understanding.

Skeleton-Based Action Segmentation: For action segmenta-

TABLE 8: Comparisons of transferred action recognition results on the PKU-MMD II.

Method	Transfer to PKU-MMD II	
	NTU-60	NTU-120
MS^{2}L [55]	45.8	-
ISC [77]	45.9	-
SCD-Net [67]	56.3	-
HiCo [63]	56.3	55.4
CMD [22]	56.0	57.0
UmURL [38]	58.2	57.6
USDRL	57.2	**58.3**

TABLE 9: Comparisons of transferred action retrieval results on the PKU-MMD II.

Method	Transfer to PKU-MMD II			
	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-set
UmURL [38]	43.4	42.8	42.5	43.0
USDRL	44.4	44.7	44.0	43.8

tion, we evaluate performance using the following metrics: frame-wise accuracy (Acc), segmental edit score (Edit), and segmental F1 scores computed at intersection-over-union thresholds of 0.10, 0.25, and 0.5 (denoted as F1@10, F1@25, and F1@50, respectively). Results on the cross-subject evaluation of the PKU-MMD II dataset are summarized in Table 7. Our approach outperforms the DeST [50] by 1.1% and 1.2% for the Acc and Edit metrics, respectively. Although our approach is not specifically designed for skeleton-based action segmentation, its performance on this task is quite competitive.

4.5 Results of Transferred Prediction Tasks

Transferred prediction aims to assess the model's capability to generalize across different datasets for the purpose of action understanding. We evaluate our approach on two tasks: transfer learning for action recognition and transfer learning for action retrieval.

Transfer Learning for Action Recognition: We assess the generalizability of the learned representations by transferring knowledge from a source dataset to a target one. Specifically, we adhere to the same experimental settings as previous studies [22], [38], [55], employing NTU-60 and NTU-120 as source datasets and PKU-MMD II as the target dataset. The evaluation is conducted following the x-sub protocol. As shown in Table 8, our proposed method attains an accuracy of 57.2% when transferring knowledge from NTU-60 and 58.3% when transferring from NTU-120, surpassing other competing methods. These results indicate that our framework effectively learns skeleton representations that generalize well to unseen datasets, a quality that is vital for real-world applications.

Transfer Learning for Action Retrieval: We setup the benchmark of transfer learning for action retrieval. We use the self-supervised models trained on the NTU-60 and NTU-120 datasets directly for action retrieval on the PKU-MMD II dataset on Table 9. Our model that is trained with the single joint modality consistently performs better than the multimodal model that is trained with joint, bone and motion modalities. These experimental results further demonstrate the generalization ability of our model.

TABLE 10: Ablation studies on the *VAC*, *XC*, *MG-FD*, and the DSTE encoder under the x-sub evaluation on the NTU-60.

VAC	*XC*	*MG-FD*	Encoder	Recog.	Retrieval
✗	✗	✗	STTR	74.8	63.8
✓	✗	✓	STTR	83.7	73.2
✓	✓	✓	STTR	**84.2**	**73.7**
✗	✗	✗	DSTE	77.3	64.4
✓	✗	✓	DSTE	84.8	74.5
✓	✓	✗	DSTE	84.6	74.7
✓	✓	✓	DSTE	**85.2**	**75.0**

4.6 Ablations, Parameter Analysis and Visualizations

Effectiveness of the VAC, XC and MG-FD: Multi-Grained Feature Decorrelation (MG-FD) consists of Variance and Auto-Covariance (VAC) term and Cross-Correlation (XC) term, and each term is regularized on multi-grained representations (MG-FD). Table 10 presents detailed ablation studies on the proposed method. Without VAC, XC and MG-FD denotes the baseline using traditional negative-based pretraining. We find that the results obtained using traditional negative-based pretraining strategies are significantly lower than those achieved with the feature decorrelation method. This demonstrates the strong potential and superiority of the feature decorrelation. Additionally, for both recognition and retrieval tasks, the method with *XC* matrix improves 0.5% over that without *XC* matrix. This indicates that the *XC* matrix could effectively capture and reduce the feature correlation between different augmented samples and help to learn more effective and robust representations. Combined with MG-FD, our approach further enhances the model's generalization ability. This improvement is attributed to the utilization of fine-grained features in the temporal, spatial and instance domains, as opposed to relying solely on the instance domain.

To further analyze the effectiveness of the proposed self-supervised representation learning paradigm, we visualize and compare learned features of different methods. As shown in Figure 5, visualizations demonstrates distinct clustering of learned features of the proposed MG-FD, suggesting that they are capable of effectively separating different classes. These results imply that the learned features capture the intrinsic characteristics of each class, enabling clear discrimination among various action categories.

Effectiveness of the CA and DSA: The proposed DSTE backbone consists of the DSA and CA modules. We evaluate the influence of the weight hyperparameters α in Eq. 5 in the DSTE structure. As shown in Figure 6, α takes values in the range [0, 1] with an interval of 0.1, where 0 and 1 represent the isolated modules of DSA and CA, respectively. We observe that both DSA and CA produce significantly inferior results compared to other settings. These results underscore the efficacy and complementarity of CA and DSA modules.

Impact of the Number of Encoder Layers: We investigate the impact of the number of encoder layers on the performance

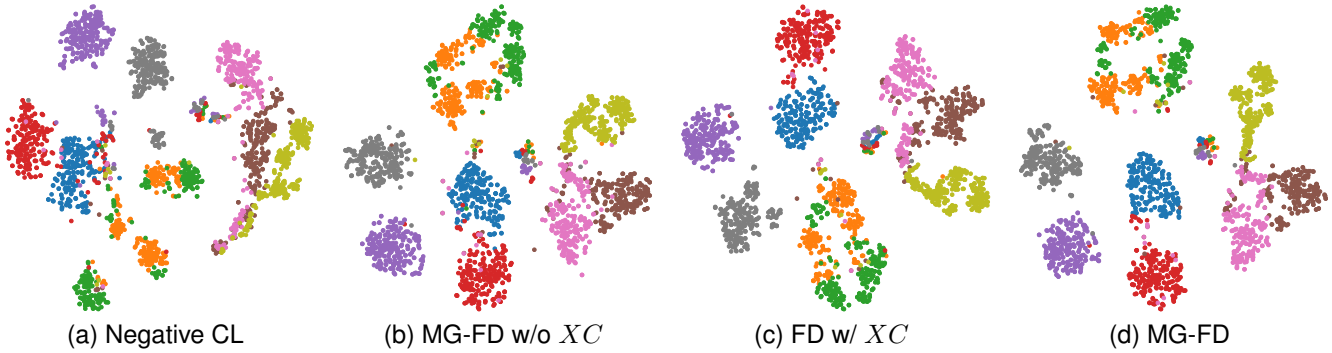


Fig. 5: Visualizations of learned instance-level representations obtained by (a) Contrastive Learning (CL), (b) Multi-Grained Feature Decorrelation (MG-FD) w/o XC, (c) Single-Grained FD w/ XC, and (d) MG-FD on the NTU-60. Nine classes from the testing set are randomly selected, and dots of the same color represent actions of the same class. We find that the model's ability to learn features that distinguish between different classes declines if multi-grained modeling or the cross-correlation (XC) term are omitted.

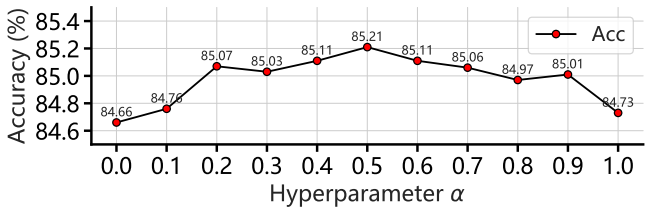


Fig. 6: The impact of weight hyperparameter α for action recognition on the xsub evaluation of the NTU-60 dataset.

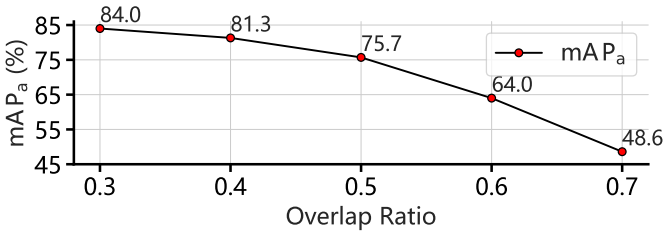


Fig. 7: Mean Average Precision (mAP$_a$) under different overlap ratios for action detection, evaluated on the PKU-MMD I dataset.

in Table 11. As the number of layers increases, the complexity of the model increases linearly. Experimental results suggest that a configuration with two layers of DSTE achieves an optimal balance between performance and complexity.

Comparison of Computation Costs: Table 12 compares the computational complexity and recognition accuracy of state-of-the-art skeleton-based action recognition models. It can be seen that our model can achieve the highest recognition accuracy with a very small amount of computational cost. For example, our model using single joint modality slightly outperforms the 3s-CMD [22], while the computational cost of ours is only 10% of that of this method.

TABLE 11: The study on the impact of the number of encoder layers. The performance is evaluated on the NTU60 X-sub dataset under the linear evaluation protocol.

Layer Num	Para. (M)	Recognition	Retrieval
1	67.7	84.5	73.1
2	94.9	85.2	75.0
3	113.9	84.9	75.4

TABLE 12: Comparison of computation costs among state-of-the-art skeleton-based action recognition methods. The performance is evaluated on the NTU60 X-sub dataset.

Method	Modality	FLOPs/G	Recognition
3s-PSTL [61]	J+M+B	3.45	79.1
3s-CrosSCLR [84]	J+M+B	17.28	82.1
3s-CMD [22]	J+M+B	17.28	84.1
UmURL [38]	J+M+B	2.54	84.2
USDRL (STTR)	J	1.74	84.2
USDRL (STTR)	J+M+B	2.54	85.8

Results of Action Detection with Different IoUs: We evaluate the performance across different overlap ratios, ranging from 0.3 to 0.7, as shown in Figure 7. Our approach achieves an mAP$_a$ of 48.6% at an overlap ratio of 0.3 and 75.7% at an overlap ratio of 0.5. These results demonstrate the efficacy of our method in dense prediction tasks by capturing fine-grained features. The robust performance across various overlap thresholds highlights our approach's adaptability, particularly in challenging scenarios. By excelling in both lower and higher overlap ratios, our model proves to be versatile, effectively adapting to different levels of temporal and spatial granularity required for accurate action detection in complex video sequences.

Visualization of Action Detection: To further demonstrate the superior performance of our approach, along with the proposed backbone DSTE, over STTR [76] in dense prediction tasks, we select a subset of video results, visualizing the action detection as depicted in Figure 8. Our approach with the DSTE not only detects a greater number of action instances but also achieves more precise localization, as indicated by the higher Intersection over Union (IoU) scores.

Analysis of Performance for Action Prediction: Given that action prediction serves as a representative dense prediction task with significant real-world applications, we analyze and compare the accuracies across different actions in Figure 9.

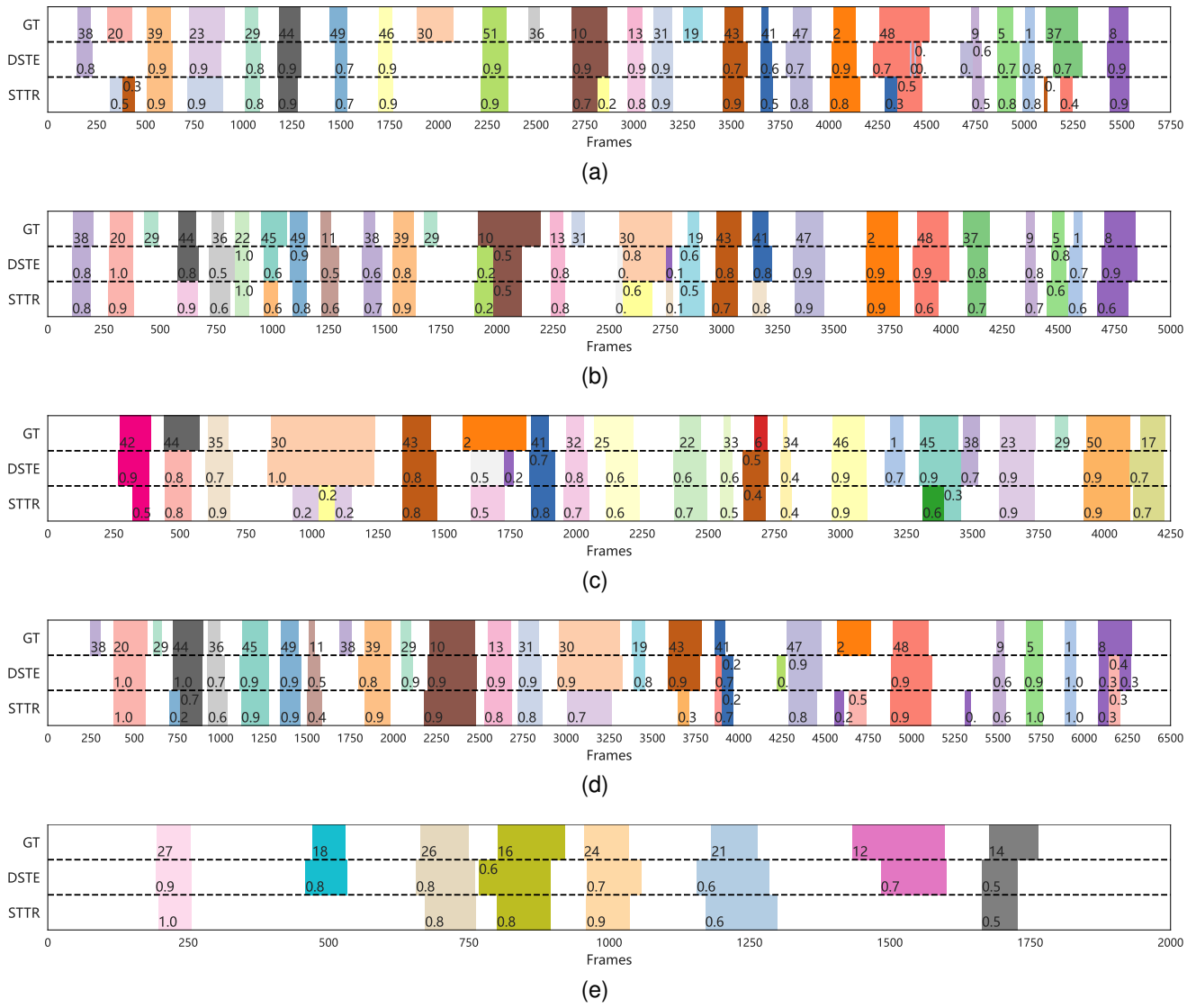
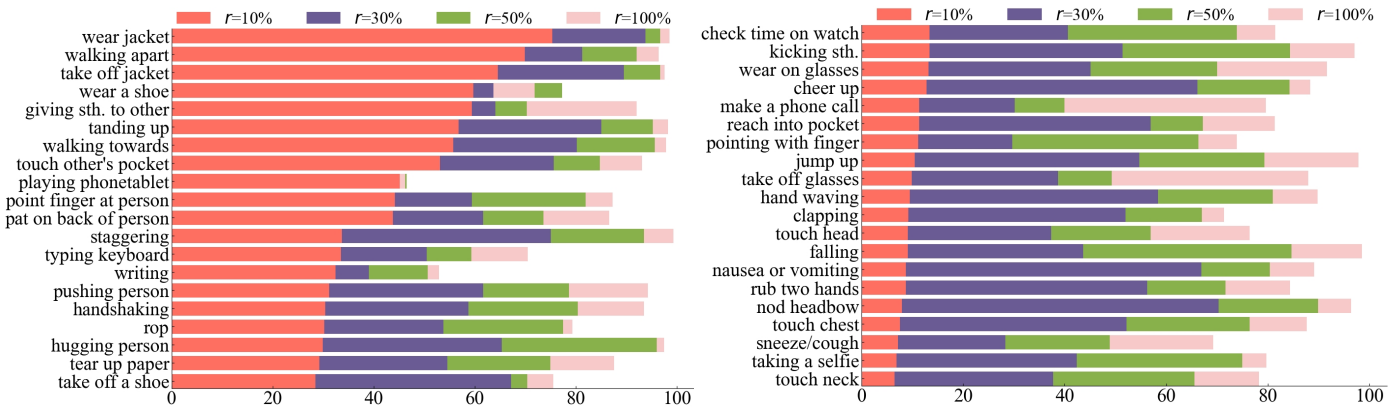


Fig. 8: Visualization of action detection in video sequences on the PKUMMD dataset. We provide results of our approach utilizing both the proposed DSTE backbone and the STTR [76] backbone. The number on the ground-truth segments denotes the index of the corresponding action class, whereas the numbers on the predicted segments represent the Intersection over Union (IoU) scores, where higher scores imply more accurate detections. Each sequence is visualized by colored segments, with each unique color distinguishing a distinct action.

We observe that certain actions, such as wear jacket, walking apart, and take off jacket, can be accurately predicted at an early stage, specifically within the first 10% of the observed sequence. When 50% of the sequence is observed, most actions can be recognized early with high accuracy, and even the more challenging actions achieve accuracies exceeding 60%, with only a few exceptions. This demonstrates the potential of our method for early action recognition in applications like condition monitoring and human–robot interaction.

5 Conclusion

We introduce a transformer-based foundation model for skeleton-based human action understanding. The proposed model comprises a Dense Spatio-Temporal Encoder (DSTE), Multi-Grained Feature Decorrelation (MG-FD), and Multi-Perspective Consistency Training (MPCT). We categorize skeleton-based human action understanding tasks into three types: coarse prediction, dense prediction, and transferred prediction, noting that dense prediction tasks have been largely overlooked. We conduct extensive experiments across eight tasks spanning the three categories and demonstrate that the proposed method serves as a unified and strong baseline for addressing them. Detailed ablation studies and analyses demonstrate the effectiveness of each module, highlighting the proposed method's high performance, strong generalization capability, and low computational cost. We believe this work could broaden the scope of research in skeleton-based action understanding and facilitate progress in dense prediction tasks, including action detection, action segmentation, and action prediction. A limitation of our approach is the need to fine-tune the linear classifier when adapting to new action understanding tasks, which limits its applicability in open-set scenarios. In the future, we aim to



(a) Top 20 Easy Categories (b) Top 20 Difficult Categories

Fig. 9: Analysis of recognition accuracy for different actions in skeleton-based action prediction on the NTU-60 dataset. r denotes the observation ratio, while the horizontal axis represents the recognition accuracy. We select the top easy and difficult action categories in (a) and (b), respectively. As the observation ratio increases, the accuracy of action recognition improves. With only the first 30% of the observation, our method can accurately recognize most actions; even for the hard categories, the accuracy reaches 40% (while random guessing accuracy is 1.6%).

improve the zero-shot generalization ability of our approach.

REFERENCES

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

[2] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[3] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[4] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 499–508.

[5] ——, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4382–4394, 2018.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[7] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[8] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192.

[9] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.

[10] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2022.

[11] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, and L. Lin, "Skeletonmae: graph-based masked autoencoder for skeleton sequence pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5606–5618.

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

[13] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[14] H. Zhang, Y. Hou, W. Zhang, and W. Li, "Contrastive positive mining for unsupervised 3d action representation learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 36–51.

[15] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 762–770.

[16] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "SSNet: Scale selection network for online 3d action prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8349–8358.

[17] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[18] W. Weng, H. Wang, J. Wang, L. He, and G.-S. Xie, "USDRL: Unified skeleton-based dense representation learning with multi-grained feature decorrelation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8332–8340.

[19] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 085–15 099.

[20] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4741–4750.

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

[22] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 734–752.

[23] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3427–3435.

[24] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins:

Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

[25] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations*, 2022.

[26] Y. Chen, L. Zhao, J. Yuan, Y. Tian, Z. Xia, S. Geng, L. Han, and D. N. Metaxas, "Hierarchically self-supervised transformer for human skeleton representation learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 185–202.

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

[28] J. Sun, L. Huang, H. Wang, C. Zheng, J. Qiu, M. T. Islam, E. Xie, B. Zhou, L. Xing, A. Chandrasekaran *et al.*, "Localization and recognition of human action in 3d using transformers," *Communications Engineering*, vol. 3, no. 1, p. 125, 2024.

[29] L. Xu, Q. Wang, X. Lin, and L. Yuan, "An efficient framework for few-shot skeleton-based temporal action segmentation," *Computer Vision and Image Understanding*, vol. 232, p. 103707, 2023.

[30] Y.-H. Li, K.-Y. Liu, S.-L. Liu, L. Feng, and H. Qiao, "Involving distinguished temporal graph convolutional networks for skeleton-based temporal action segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 647–660, 2023.

[31] D. Yang, Y. Wang, A. Dantcheva, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond, "Lac-latent action composition for skeleton-based action segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 679–13 690.

[32] S. W. Hyder, M. Usama, A. Zafar, M. Naufil, F. J. Fateh, A. Konin, M. Z. Zia, and Q.-H. Tran, "Action segmentation using 2d skeleton heatmaps and multi-modality fusion," in *IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 1048–1055.

[33] H. Ji, B. Chen, X. Xu, W. Ren, Z. Wang, and H. Liu, "Language-assisted skeleton action understanding for skeleton-based temporal action segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 400–417.

[34] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," in *IEEE International Conference on Computer Vision*, vol. 1, no. 2, 2017.

[35] Y. Kong, S. Gao, B. Sun, and Y. Fu, "Action prediction from videos via memorizing hard-to-predict samples," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[36] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3556–3565.

[37] X. Wang, Z. Fang, X. Li, X. Li, C. Chen, and M. Liu, "Skeleton-in-context: Unified skeleton sequence modeling with in-context learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2436–2446.

[38] S. Sun, D. Liu, J. Dong, X. Qu, J. Gao, X. Yang, X. Wang, and M. Wang, "Unified multi-modal unsupervised representation learning for skeleton-based action understanding," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 2973–2984.

[39] L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, "Unified pose sequence modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 019–13 030.

[40] H. Qu, Y. Cai, and J. Liu, "LLMs are good action recognizers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 395–18 406.

[41] H. Wang, X. Ma, J. Kuang, and J. Gui, "Heterogeneous skeleton-based action representation learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 154–19 164.

[42] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3015–3024.

[43] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6787–6800.

[44] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[45] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5036–5045.

[46] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 549–14 560.

[47] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bevt: Bert pretraining of video transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 733–14 743.

[48] Y. Xu and Y. Fu, "Sports-Traj: A unified trajectory generation model for multi-agent movement in sports," in *The Thirteenth International Conference on Learning Representations*, 2025.

[49] M. Zhang, D. Jin, C. Gu, F. Hong, Z. Cai, J. Huang, C. Zhang, X. Guo, L. Yang, Y. He *et al.*, "Large motion model for unified multi-modal motion generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 397–421.

[50] Y. Li, Z. Li, S. Gao, Q. Wang, Q. Hou, and M.-M. Cheng, "A decoupled spatio-temporal framework for skeleton-based action segmentation," *arXiv preprint arXiv:2312.05830*, 2023.

[51] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[52] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[53] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–24, 2020.

[54] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 266–16 275.

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

[56] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3d action representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 423–13 433.

[57] B. Kim, H. J. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 209–225.

[58] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, and A. Lu, "Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition," in *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2023, pp. 224–229.

[59] L. Wu, L. Lin, J. Zhang, Y. Ma, and J. Liu, "MacDiff: Unified skeleton modeling with masked conditional diffusion," in *European Conference on Computer Vision*. Springer, 2024, pp. 110–128.

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

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

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

[63] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang, "Hierarchical contrast for unsupervised skeleton-based action representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[64] S. Guan, X. Yu, W. Huang, G. Fang, and H. Lu, "Dmmg: dual min-max games for self-supervised skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 395–407, 2023.

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

[66] X. Wang and Y. Mu, "Localized linear temporal dynamics for self-supervised skeleton action recognition," *IEEE Transactions on Multimedia*, 2024.

[67] C. Wu, X.-J. Wu, J. Kittler, T. Xu, S. Ahmed, M. Awais, and Z. Feng, "Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5949–5957.

[68] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Brémond, "View-invariant skeleton action representation learning via motion retargeting," *International Journal of Computer Vision*, pp. 1–16, 2024.

[69] L. Lin, L. Wu, J. Zhang, and J. Liu, "Idempotent unsupervised representation learning for skeleton-based action recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 75–92.

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

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

[72] Z. Jin, Y. Wang, Q. Wang, Y. Shen, and H. Meng, "Ssrl: Self-supervised spatial-temporal representation learning for 3d action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 274–285, 2023.

[73] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2363–2372.

[74] Y. Zhu, H. Han, Z. Yu, and G. Liu, "Modeling the relative visual tempo for self-supervised skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 913–13 922.

[75] M. Wang, X. Li, S. Chen, X. Zhang, L. Ma, and Y. Zhang, "Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 3207–3220, 2023.

[76] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

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

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

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

[80] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1473–1481.

[81] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang, "Early action prediction by soft regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2568–2583, 2018.

[82] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.

[83] Y. Li, Z. Dong, K. Liu, L. Feng, L. Hu, J. Zhu, L. Xu, S. Liu *et al.*, "Efficient two-step networks for temporal action segmentation," *Neurocomputing*, vol. 454, pp. 373–381, 2021.

[84] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4741–4750.