Hierarchical-aware Orthogonal Disentanglement Framework for Fine-grained Skeleton-based Action Recognition

Haochen Chang[1], Pengfei Ren[2*], Haoyang Zhang[3,4,5], Liang Xie[3,4,5], Hongbo Chen[1*], Erwei Yin[3,4,5]
[1] School of Systems Science and Engineering, Sun Yat-sen University
[2] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[3] Defense Innovation Institute, Academy of Military Sciences
[4] Intelligent Game and Decision Laboratory [5] Tianjin Artificial Intelligence Innovation Center
changhch5@mail2.sysu.edu.cn, rpf@bupt.edu.cn, haoyang@tju.edu.cn
{xielnudt,yinerwei1985}@gmail.com, chenhongbo@mail.sysu.edu.cn

Abstract

In recent years, skeleton-based action recognition has gained significant attention due to its robustness in varying environmental conditions. However, most existing methods struggle to distinguish fine-grained actions due to subtle motion features, minimal inter-class variation, and they often fail to consider the underlying similarity relationships between action classes. To address these limitations, we propose a Hierarchical-aware Orthogonal Disentanglement framework (HiOD). We disentangle coarse-grained and fine-grained features by employing independent spatial-temporal granularity-aware bases, which encode semantic representations at varying levels of granularity. Additionally, we design a cross-granularity feature interaction mechanism that leverages complementary information between coarse-grained and fine-grained features. We further enhance the learning process through hierarchical prototype contrastive learning, which utilizes the parent class hierarchy to guide the learning of coarse-grained features while ensuring the distinguishability of fine-grained features within child classes. Extensive experiments on FineGYM, FSD-10, NTU RGB+D, and NTU RGB+D 120 datasets demonstrate the superiority of our method in fine-grained action recognition tasks.

1. Introduction

Due to advancements in pose estimation algorithms [1, 10, 24, 26, 30, 37], action recognition has seen widespread application in human–computer interaction in recent years [13, 17, 25, 27]. Skeleton data has garnered significant attention from researchers due to its robustness to variations in viewpoint and illumination. Early skeleton-based action recognition approaches primarily utilized RNNs [31, 34, 44] and CNNs [15] to process skeleton data as pseudo-images or feature sequences, which overlooked the inherent correlations between joints. Recent approaches [5, 7, 14, 22, 28, 33, 35, 38, 41] employ Graph Convolutional Networks (GCNs) to model the skeleton as a graph, leveraging the topological structure of the skeletons to enhance recognition accuracy.

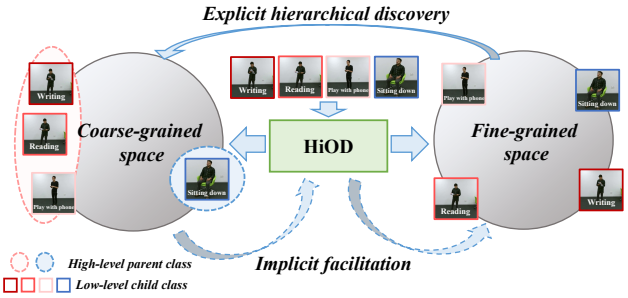

Figure 1. HiOD uses fine-grained feature clustering to establish explicit parent-child hierarchies, guiding coarse-grained feature learning and aiding in finding parent class boundaries. Parent class classification priors provide implicit disambiguation cues for fine-grained features via shared bases, refining child class distinctions.

However, most existing methods struggle to identify fine-grained actions due to subtle motion features, minimal inter-class variation, and large intra-class variability. Recently, some methods extract local detailed features through multi-level feature aggregation [23], attention mechanisms [4, 11, 19], or carefully designed loss functions [3, 21, 42]. Nevertheless, these methods overlook the interference that coarse-grained features can have on fine-grained feature learning. Specifically, in action recognition tasks, most

*Corresponding Author.

samples exhibit significant inter-class differences and can be distinguished by certain salient features known as coarse-grained features. Optimizing easily distinguishable classes can rapidly decrease the overall classification error, resulting in coarse-grained features dominating the learning process and causing fine-grained features to be overwhelmed.

Furthermore, fine-grained actions exhibit complex interrelationships and possess a hierarchical semantic parent-child structure [16]. For instance, compared to "Sit down" class, "Reading" class exhibits more similar spatiotemporal representations to "Writing" class, suggesting they may belong to the same semantic parent class, as shown in Figure 1. High-level parent classes can be distinguished using coarse-grained features, while low-level child classes rely on fine-grained features. Fine-grained features can provide detailed supplementary information to coarse-grained features, facilitating the mining of general patterns shared among child classes. Coarse-grained features contain prior information about high-level classes, which provide significant disambiguation cues for the learning of fine-grained features. Therefore, we aim to leverage the class hierarchy to mitigate the adverse effects of coarse-grained features on fine-grained features in deep semantic space, while also learning intrinsic associations between these features.

Motivated by the above, we propose a Hierarchical-aware Orthogonal Disentanglement framework (HiOD) for fine-grained skeleton-based action recognition. Specifically, HiOD consists of two core parts: Granularity Disentanglement (GrD) and Hierarchical Prototype Contrastive Learning (HiPCL). GrD employs independent spatial-temporal granularity-aware bases to encode semantic representations at varying levels of granularity, thereby disentangling coarse- and fine-grained features into distinct subspaces. To achieve cross-granularity feature interaction, GrD employs a shared orthogonal basis to project disentangled features into a unified semantic space, capturing their common variation patterns and uncovering the complementary information between them. To further promote the disentanglement and interaction of features, HiPCL utilizes a soft prototype weighted contrastive loss based on a clustering hierarchy to capture shared semantic representations among similar classes, dynamically construct high-level parent class decision boundaries, and achieve more robust coarse-grained feature learning. In contrast, HiPCL employs a hard prototype contrastive loss to ensure the semantic distinguishability of subclasses, thereby achieving discriminative fine-grained feature learning.

The main contributions are summarized as follows:

- We propose a hierarchical-aware orthogonal disentanglement framework that uses spatial-temporal granularity-aware bases and shared orthogonal bases to mitigate the negative impact of coarse-grained features on fine-grained features while learning their intrinsic associations.
- We propose a hierarchical prototype contrastive learning approach that exploits the parent class hierarchy to guide the learning of coarse-grained features while ensuring the child class distinguishability of fine-grained features.
- We conduct extensive experiments on FineGYM, FSD-10, NTU RGB+D, and NTU RGB+D 120 datasets to compare our proposed methods with the state-of-the-art models. Results demonstrate the significant improvement achieved by our methods in fine-grained action recognition.

2. Related Work

Skeleton-based action recognition. Early deep learning methods employed RNNs [31, 34, 44] and CNNs [15] to extract action features from skeleton data. However, both fail to effectively capture the structured dependencies of skeletons due to their inherent calculation strategies. To address these challenges, some methods [5, 14, 38] model the human skeleton as a graph to capture the skeletal topology. However, these methods struggle to capture the subtle differences between similar actions. Recently, some methods focused on fine-grained skeleton-based action recognition [3, 11, 21, 23, 42]. Zhou *et al*. [42] decoupled spatial-temporal features and used contrastive learning to refine discriminative representation. Geng *et al*. [11] proposed a self-attention-enhanced graph neural network integrating angle information for richer skeleton features. Liu *et al*. [21] developed channel-variable spatial-temporal attention and a robust decoupled loss to boost joint discrimination. However, these methods ignore the influence of coarse-grained features on fine-grained features during training. We aim to mitigate the negative impact of coarse-grained features on fine-grained features while fully exploring intrinsic associations for complementary benefits.

Disentangled representation learning. Disentangled representation learning separates factors of variation, enhancing interpretability and generalization across various tasks. Current methods achieve feature disentanglement through auxiliary tasks, including motion-context decoupling [12], reconstruction-based separation [6, 29], and temporal-invariant decomposition [45]. Inspired by [39, 40] and recognizing the highly semantic nature of human actions, we utilize orthogonal bases to facilitate the disentanglement and interaction of coarse-grained and fine-grained features. Notably, we use hierarchical prototype contrastive learning to guide disentanglement without auxiliary tasks.

Hierarchical classification. Hierarchical classification uses hierarchical multi-labels to provide rich supervision to the network. The challenge is embedding label hierarchy into loss functions or architectures. Existing methods embed hierarchies via loss reweighting [2] or contrastive space enhancement [16], but require predefined parent-child relations. For skeleton actions without explicit hierarchies, we guide disentanglement of coarse- and fine-grained features

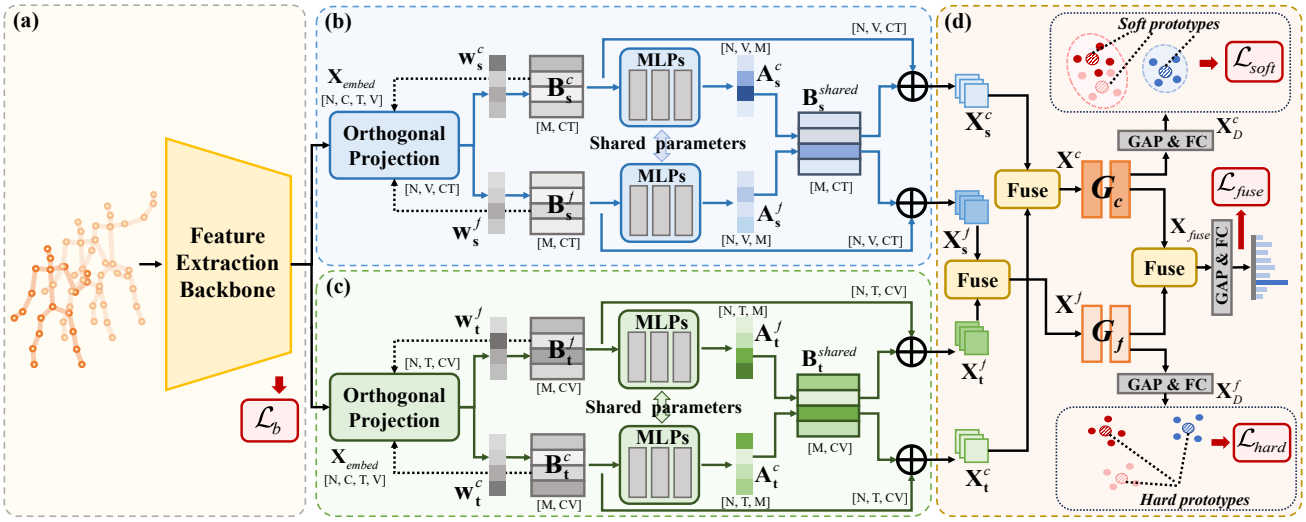

Figure 2. **Overview of our proposed HiOD. (a)** Feature extraction backbone embeds skeleton action features into high-level semantic space. **(b)∼(c)** GrD in spatial and temporal dimensions. The granularity-aware bases $\mathbf{B}_\mathbf{s}^c$, $\mathbf{B}_\mathbf{s}^f$, $\mathbf{B}_\mathbf{t}^c$, and $\mathbf{B}_\mathbf{t}^f$ capture semantic representations across different granularities. Meanwhile, the shared orthogonal bases $\mathbf{B}_\mathbf{s}^{shared}$ and $\mathbf{B}_\mathbf{t}^{shared}$ extract complementary information from the disentangled features within the same semantic space. **(d)** HiPCL leverages the parent-child hierarchy to direct the learning of coarse-grained features while maintaining the discriminability of fine-grained features through high-confidence prototypes.

via dual weighting of prototype updates and contrastive loss, revealing parent-child relationships among classes.

3. Method

3.1. Overview

The overall architecture of our proposed Hierarchical-aware Orthogonal Disentanglement framework (HiOD) is depicted in Figure 2. Initially, the feature extraction backbone maps the skeleton features into embedding space to obtain the high-level embedding feature $\mathbf{X}_{embed} \in \mathbb{R}^{N\times C\times T\times V}$, where N is the batch size, C is the output channel dimensions, T is the number of frames, and V is the number of joints. Next, Granularity Disentanglement (GrD) projects the embedding features into subspaces of different granularities in both temporal and spatial dimensions to disentangle fine-grained features from coarse-grained features. Concurrently, these disentangled features are projected into a common semantic space to facilitate cross-granularity interaction. Finally, coarse-grained and fine-grained features are fused for action classification. In this process, Hierarchical Prototype Contrastive Learning (HiPCL) leverages the parent-child hierarchy to guide the learning of coarse-grained features while maintaining the discriminability of fine-grained features through high-confidence prototypes. We present the architecture details in the following sections.

3.2. Granularity disentanglement

Granularity-aware basis. To mitigate the adverse effects of coarse-grained features on fine-grained features, we project the original action features $\mathbf{X}_{embed}$ into a subspace with specific granularity using a parameterized granularity-aware orthogonal basis. Formally, the granularity-aware basis $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M\}$ consists of M orthogonal vectors, with each $\mathbf{b}_M \in \mathbb{R}^{1\times C_{dim}}$ representing a semantic granularity. We implement $\mathbf{B} \in \mathbb{R}^{M\times C_{dim}}$ as a learnable matrix and employ QR decomposition during the forward pass to enforce orthogonality:

$$\mathbf{b}_i \cdot \mathbf{b}_j = \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases} \tag{1}$$

In other words,we enforce orthogonality among vectors within and across orthogonal bases. Assuming that the embedding feature can be represented as a vector cluster $\mathbf{X}_v$, we first calculate the projection coefficients $\mathbf{w} = [w_1, w_2, \ldots, w_i]$ of $\mathbf{X}_v$ onto each basis vector. Next, we perform a weighted sum of all coefficients with their corresponding basis vectors. This allows us to project the original skeleton embedding features $\mathbf{X}_{embed}$ onto the subspace spanned by the granularity-aware basis. The process can be expressed as:

$$\mathbf{X}_b = \sum_{i=1}^{M} w_i \mathbf{b}_i^T,\ w_i = \frac{\mathbf{X}_v \cdot \mathbf{b}_i^T}{\|\mathbf{b}_i\|^2},\quad i = 1, 2, \ldots, M \tag{2}$$

To disentangle coarse-grained and fine-grained features, we employ two independent granularity-aware bases and guide the learning process of these orthogonal bases using hierarchical prototype contrastive learning (see Section 3.3). Each basis vector encodes semantic representations

of different granularities and adaptively selects coarse- and fine-grained information from the original features via feature projection, creating feature subspaces of varying granularities.Therefore, the granularity-aware basis also functions as a feature selector.

Spatio-temporal separation strategy. Subtle differences among various fine-grained actions can manifest across different spatial-temporal dimensions. For example, in rhythmic gymnastics, jumps can be classified based on variations in body posture (spatial dimension) and the number of rotations (temporal dimension). Thus, separately disentangling coarse-grained and fine-grained features in both temporal and spatial dimensions is beneficial. Specifically, we design four independent granularity-aware bases $\mathbf{B}_\mathbf{t}^c$, $\mathbf{B}_\mathbf{t}^f$, $\mathbf{B}_\mathbf{s}^c$, $\mathbf{B}_\mathbf{s}^f$ to capture temporal coarse-grained features, temporal fine-grained features, spatial coarse-grained features, and spatial fine-grained features. We reshape the embedding feature $\mathbf{X}_{embed} \in \mathbb{R}^{N\times C\times T\times V}$ into spatial feature $\mathbf{X}_\mathbf{s} \in \mathbb{R}^{N\times V\times CT}$ and temporal feature $\mathbf{X}_\mathbf{t} \in \mathbb{R}^{N\times T\times CV}$, and project them into the subspaces defined by the four granularity-aware bases, respectively:

$$\begin{aligned} \mathbf{X}_\mathbf{s}^c &= (\mathbf{X}_\mathbf{s}\cdot(\mathbf{B}_\mathbf{s}^c)^T)\cdot\mathbf{B}_\mathbf{s}^c, \quad \mathbf{X}_\mathbf{s}^f = (\mathbf{X}_\mathbf{s}\cdot(\mathbf{B}_\mathbf{s}^f)^T)\cdot\mathbf{B}_\mathbf{s}^f, \\ \mathbf{X}_\mathbf{t}^c &= (\mathbf{X}_\mathbf{t}\cdot(\mathbf{B}_\mathbf{t}^c)^T)\cdot\mathbf{B}_\mathbf{t}^c, \quad \mathbf{X}_\mathbf{t}^f = (\mathbf{X}_\mathbf{t}\cdot(\mathbf{B}_\mathbf{t}^f)^T)\cdot\mathbf{B}_\mathbf{t}^f. \end{aligned} \tag{3}$$

where c and f represent the coarse-grained and fine-grained, respectively. Eq.3 represents the matrix multiplication implementation of the orthogonal projection (Eq.2). In this way, granularity-aware bases to encode specific temporal and spatial dynamics. The temporal coarse-grained basis captures overarching motion patterns over long time intervals, allowing the model to understand general action trends. In contrast, the temporal fine-grained basis focuses on minute variations in shorter time frames. Similarly, the spatial coarse-grained basis emphasizes broad spatial configurations of joints to detect larger motion patterns, while the spatial fine-grained basis captures precise movements and interactions among individual joints.

Cross-granularity feature interaction. Some methods adhere to the principle of completely disentangle features by utilizing separate network heads for feature separation [2, 6]. However, this approach results in a lack of correlation between the heads, which makes it challenging to model the implicit relationships among different features. In contrast, we aim to effectively leverage the complementary information between coarse-grained and fine-grained features to promote each other while ensuring their separation to the greatest extent possible. To this end, we propose a cross-granularity feature interaction mechanism that projects coarse-grained and fine-grained features into the same semantic subspace, enabling the extraction of complementary information between them. We adopt two learnable shared orthogonal bases, $\mathbf{B}_\mathbf{t}^{shared} \in \mathbb{R}^{M\times CV}$ and $\mathbf{B}_\mathbf{s}^{shared} \in \mathbb{R}^{M\times CT}$, to facilitate the interaction of features with different granularities in the temporal and spatial dimensions, respectively. We take the spatial dimension as an example to introduce the specific technical details. We project coarse-grained features $\mathbf{X}_\mathbf{s}^c \in \mathbb{R}^{N\times V\times CT}$ and fine-grained features $\mathbf{X}_\mathbf{s}^f \in \mathbb{R}^{N\times V\times CT}$ into the same subspace using a spatial-shared orthogonal basis $\mathbf{B}_\mathbf{s}^{shared}$, which serves as a prerequisite for effective interaction. Unlike linear orthogonal projection, the projection coefficients $\mathbf{A}_\mathbf{s}^c$ and $\mathbf{A}_\mathbf{s}^f$ of the shared orthogonal basis $\mathbf{B}_\mathbf{s}^{shared}$ are learned through a $\mathbf{MLP}$ layer. This ensures the dynamism and adaptability of cross-granularity interactions while alleviating the strong constraints of orthogonal disentanglement. Finally, we utilize the projection coefficients to perform a weighted summation of the shared basis vectors, resulting in the cross-granularity supplementary features $\mathbf{A}_\mathbf{s}^c \cdot \mathbf{B}_\mathbf{s}^{shared}$ and $\mathbf{A}_\mathbf{s}^f \cdot \mathbf{B}_\mathbf{s}^{shared}$. This process can be expressed as follows:

$$\begin{aligned} \mathbf{X}_\mathbf{s}^c &= \mathbf{X}_\mathbf{s}^c + \mathbf{A}_\mathbf{s}^c \cdot \mathbf{B}_\mathbf{s}^{shared},\ \mathbf{A}_\mathbf{s}^c = \mathbf{MLP}\left(\mathbf{X}_\mathbf{s}^c\right), \\ \mathbf{X}_\mathbf{s}^f &= \mathbf{X}_\mathbf{s}^f + \mathbf{A}_\mathbf{s}^f \cdot \mathbf{B}_\mathbf{s}^{shared},\ \mathbf{A}_\mathbf{s}^f = \mathbf{MLP}\left(\mathbf{X}_\mathbf{s}^f\right). \end{aligned} \tag{4}$$

Since coarse-grained features and fine-grained features are encoded using the same orthogonal basis $\mathbf{B}_\mathbf{s}^{shared}$, $\mathbf{B}_\mathbf{s}^{shared}$ serves as a medium for interaction between the two. In particular, the shared orthogonal basis provides a common projection space for features of different granularities, allowing them to be projected and aligned on the same basis vectors. These basis vectors capture the latent correlations between features, enabling the identification of their common variation patterns and assisting the model in understanding the complementary information between them.

3.3. Hierarchical Prototype Contrastive Learning

Coarse-grained features can represent the overall patterns of actions and aid in distinguishing the action parent class, such as the jump class in rhythmic gymnastics. In contrast, fine-grained features can capture the detailed distinctions among the various child classes of the jumping class. This motivates us to utilize the class hierarchy to guide the learning of both coarse-grained and fine-grained features.

Hierarchical Discovery. In skeleton-based action recognition, most datasets typically provide only the finest-grained labels without predefined hierarchical structures, making it difficult to directly infer the hierarchical relationships among action classes. Since classes at the same level often exhibit high semantic similarity, we can leverage the similarity of embedded features between samples to uncover the hierarchical structure. Specifically, we employ the K-Means clustering algorithm to cluster the fine-grained features into K parent classes, resulting in a cluster assignment matrix $\mathbf{S} \in \mathbb{R}^{K\times N_c}$, where N_c represents the total number of classes in the dataset. The cluster assignment matrix $\mathbf{S}$ reflects the hierarchical relationships across different classes.

Soft and hard prototypes. To ensure that the parent-child hierarchical relationships guide the learning of coarse-grained and fine-grained features, we propose the concepts of hard prototypes and soft prototypes, as shown in Figure 3(a). Hard prototypes $\left\{\boldsymbol{P}_n^{hard} \in \mathbb{R}^D\right\}_{n=1}^{N_c}$ are designed to represent the distinctive fine-grained characteristics of each class, ensuring clear separability between subclasses. Conversely, soft prototypes $\left\{\boldsymbol{P}_n^{soft} \in \mathbb{R}^D\right\}_{n=1}^{N_c}$ are designed to incorporate the hierarchical relationships between classes, extracting shared semantic features that reflect broader, coarse-grained patterns. By combining hard and soft prototypes, our approach effectively captures both the fine-grained nuances and coarse-grained similarities between classes, resulting in facilitating the disentanglement of coarse-grained and fine-grained features. Specifically, we maintain these prototypes using a momentum update approach:

$$\boldsymbol{P}_n^{hard} \leftarrow (1-\alpha)\cdot\frac{\sum_{i\in s^n}\mathbb{1}\left(p_{logits}>t\right)\cdot\mathbf{X}_{D,i}^f}{\sum_{i\in s^n}\mathbb{1}\left(p_{logits}>t\right)}+\alpha\cdot\boldsymbol{P}_n^{hard}, \tag{5}$$

$$\boldsymbol{P}_n^{soft} \leftarrow (1-\alpha)\cdot\sum_{i\in\{s^n\}_{n=1}^{N_c}}\mathbf{S}_{k,n}\cdot\mathbf{X}_{D,i}^c+\alpha\cdot\boldsymbol{P}_n^{soft}, \tag{6}$$

where n represents the n-th action class. α is the momentum term, we set it to 0.9 by experience. s^n is the sample set of class n in the current batch. $\mathbb{1}(\cdot)$ is the indicator function. When the logits probability of sample i is greater than the threshold, $\mathbb{1}\left(p_{logits}>t\right)=1$, and when it is less than the threshold, $\mathbb{1}\left(p_{logits}>t\right)=0$, that is, the sample does not participate in the update of the prototype representation. $\mathbf{X}_{D,i}^c \in \mathbb{R}^D$ and $\mathbf{X}_{D,i}^f \in \mathbb{R}^D$ are the coarse-grained feature and the fine-grained feature after GAP and FC, respectively. $\mathbf{S}_{k,n}$ represents the element at the (k,n) position of the cluster assignment matrix after applying row softmax, indicating the probability that the n-th subclass belongs to the k-th parent class. This value indirectly reflects the semantic similarity between classes. To mitigate the effects of class imbalance, we normalize $\mathbf{S}_{k,n}$ by the number of samples in n-th subclass.

Hard prototypes are computed using high-confidence fine-grained features from the same class samples, which effectively reduces noise interference and enhances intra-class consistency of the prototypes. Soft prototypes are obtained by weighting the coarse-grained features of all samples in the batch, representing the inter-class similarity. It is noteworthy that soft prototypes are not prototypes for the K parent classes but are coarse-grained feature prototypes for the N_c subclasses, reflecting the high-level semantics of each class. Consequently, soft prototypes belonging to the same parent class will exhibit similarity, but they will not be identical, ensuring inter-class resolution.

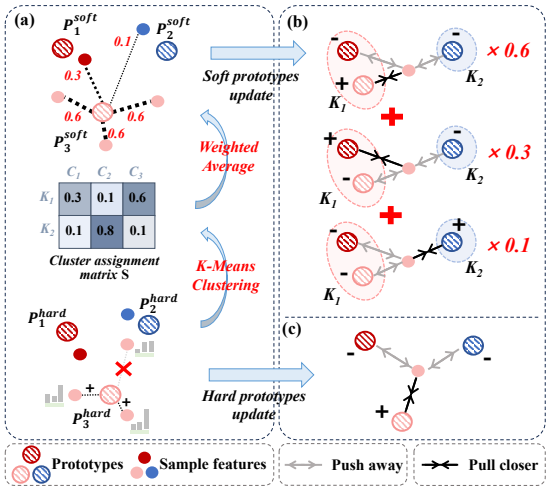

Figure 3. **Illustration of our proposed HiPCL. (a)** Hard prototypes are calculated as the average of high-confidence similar samples. While soft prototypes are obtained by a weighted summation of all class samples in the batch, using weights from the clustering assignment matrix. **(b)** Soft prototype weighted contrastive loss. **(c)** Standard prototype contrastive loss.

Soft prototype weighted contrastive loss. We employ the standard prototype contrastive loss $\mathcal{L}_{hard}$ to ensure that the fine-grained features are close to the hard prototype of their respective class while distancing themselves from the hard prototypes of other classes. In contrast, we propose a soft prototype weighted contrastive loss $\mathcal{L}_{soft}$ to facilitate the learning of coarse-grained features. $\mathcal{L}_{soft}$ considers soft prototypes from all categories and assigns greater weight to those with stronger correlations, as determined by the cluster assignment matrix. Therefore, we represent $\mathcal{L}_{soft}$ as a weighted sum of multiple standard prototype contrastive losses. $\mathcal{L}_{hard}$ and $\mathcal{L}_{soft}$ can be expressed as:

$$\mathcal{L}_{hard}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{X}_{D,i}^f\cdot\boldsymbol{P}_n^{hard}/\tau\right)}{\sum_{l\neq n}\exp\left(\mathbf{X}_{D,i}^f\cdot\boldsymbol{P}_l^{hard}/\tau\right)}, \tag{7}$$

$$\mathcal{L}_{soft}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{n=1}^{N_c}\mathbf{S}_{k,n}\log\frac{\exp\left(\mathbf{X}_{D,i}^c\cdot\boldsymbol{P}_n^{soft}/\tau\right)}{\sum_{l\neq n}\exp\left(\mathbf{X}_{D,i}^c\cdot\boldsymbol{P}_l^{soft}/\tau\right)}, \tag{8}$$

where $\mathbf{X}_i^c$ and $\mathbf{X}_i^f$ are the coarse-grained and fine-grained features, respectively. τ is the temperature coefficient. $\mathbf{S}_{k,n}$ is the clustering assignment weight obtained after applying the softmax function.

Feature refinement. The coarse-grained and fine-grained features mentioned in this section are all obtained after spatiotemporal addition fusion, as demonstrated in Figure 2(d). The coarse-grained and fine-grained features are input into the G_c and G_f blocks for further refinement. Both blocks consist of two layers of spatiotemporal feature extraction,

using the same GCN-TCN architecture as the final layer of the backbone. G_c and G_f core role is to establish independent optimization paths for feature at each granularity and learn a representation space that aids disentanglement. We will discuss the role of the refinement module in the Supplementary Material C.2.

3.4. Training Objective

To effectively utilize both coarse-grained and fine-grained features, we combine them by summing, resulting in a fused feature $\mathbf{X}_{fuse}$ that outputs the action class. We use cross-entropy loss to optimize the fusion feature $\mathbf{X}_{fuse}$. Additionally, we employ the cross-entropy loss $\mathcal{L}_b$ to supervise $\mathbf{X}_{embed}$, providing intermediate supervision to establish the initial feature space, which aids in decoupling and stabilizing training. Finally, the full learning objective function can be expressed as:

$$\mathcal{L} = \lambda_b \cdot \mathcal{L}_b + \lambda_{fuse} \cdot \mathcal{L}_{fuse} + \lambda_{hard} \cdot \mathcal{L}_{hard} + \lambda_{soft} \cdot \mathcal{L}_{soft}, \tag{9}$$

4. Experiments

4.1. Datasets

FineGYM. FineGYM [32] is a large-scale fine-grained action recognition dataset for gymnastic videos, containing 29,000 videos across 99 fine-grained gymnastics action classes. The intricate interplay of flips, twists, and holds in gymnastic actions creates significant recognition ambiguities. We follow the previous method [8, 21] to extract the skeleton data.

FSD-10. FSD-10 [20] is a fine-grained figure skating dataset with 1,484 clips, capturing 10 men/ladies' actions at 30 fps (1080×720). Each sample includes 1,500 frames and 25 skeleton joints. The subtle combinations of jumps, spins, and steps significant recognition challenges.

NTU RGB+D. NTU RGB+D [31] is a large-scale action recognition dataset with 56,800 skeleton sequences performed by 40 volunteers across 60 classes. Each sequence is captured by Microsoft Kinect v2. It offers two benchmarks: X-Sub and X-View.

NTU RGB+D 120. NTU-RGB+D 120 [18] is an extension of the original NTU RGB+D 60 dataset, incorporating an additional 57,367 action samples, resulting in a total of 113,945 skeleton action samples categorized into 120 classes, with contributions from 106 distinct subjects. It provides two benchmarks: X-Sub and X-Set.

4.2. Implementation Details

We implement our method with PyTorch framework and perform all experiments on one RTX4090D GPU. We train our models using SGD with a momentum of 0.9 and weight decay of 0.0004. The batch size is set to 64 and the base learning rate is set to 0.1. For NTU RGB+D and NTU RGB+D 120, the learning rate decays with a factor of 0.1 at epoch 35, 55 and 75 for 130 epochs. The number of clustering classes K is set to 40 and 80, respectively. For FineGYM and FSD-10, the learning rate decays with a factor of 0.1 at epoch 35, 65 and 95 for 120 epochs, with K set to 60 and 5, respectively. The hyper-parameters in our methods are set as: $\lambda_b = 0.5$, $\lambda_{fuse} = 1$, $\lambda_{hard} = 1$, $\lambda_{soft} = 0.5$.

4.3. Comparison with State-of-the-Arts Methods

Overall performance comparison. We conduct a comparison with the state of-the-art methods on FineGYM, FSD-10 and NTU RGB+D 60/120 datasets to demonstrate the competitive ability of our proposed method. Following the previous approach [14], we adopt the six-way ensemble method to report the results. Comparisons for each dataset are shown in Table 1. Our model achieves state-of-the-art performance with a large margin on the fine-grained datasets FineGYM and FSD-10. On the coarse-grained datasets NTU RGB+D 60/120 datasets, our method also outperforms the SOTA models, suggesting that our method effectively balances coarse- and fine-grained features.

Fine-grained analysis on confusing actions. To further illustrate that our method alleviates the interference of coarse-grained features on fine-grained features, we present the recognition results for some easily confusing actions in Table 2. According to the classification results of CTR-GCN [5], we partition the NTU60 dataset into three subdatasets based on difficulty: actions with accuracy below 70% are classified as hard set, those between 70% and 90% as medium set, and those above 90% as easy set. The hard set typically contains confusing actions. We compared our method with three other methods on these subdatasets. The results in the first three rows of Table 2 show that our method improves performance across all subdatasets, achieving the highest increase of 3.9% in the hard dataset. The last four rows indicate that our method improved by over 4% for all four classes of confusing actions. These results indicate that our method provides a significant advantage in distinguishing confusing action classes.

4.4. Ablation Studies

We conduct ablation studies using joint modality on the FineGYM dataset, which contains more fine-grained samples. We build the baseline with the backbone proposed in [14] and train it using cross-entropy loss. In addition, we provide more experimental results in supplementary materials.

Effectiveness of granularity disentanglement. To verify the effectiveness of granularity disentanglement module, we gradually add the components to baseline model to conduct comparative experiments. Specifically, we set up four model variants: (1) using only spatial granularity-aware bases; (2) using only temporal granularity-aware bases; (3) using both spatial and temporal granularity-aware bases;

Method	FineGYM	FSD-10	NTU RGB+D X-Sub	NTU RGB+D X-View	NTU RGB+D 120 X-Sub	NTU RGB+D 120 X-Set
ST-GCN [38]	86.7	84.2	81.5	88.3	-	-
2s-AGCN [33]	-	88.2	88.5	95.1	82.9	84.9
MS-G3D [22]	92.0	88.7	91.5	96.2	86.9	88.4
MS-G3D++ [9]	92.6	-	92.2	96.6	87.2	89.0
CTR-GCN [5]	91.9	90.6	92.4	96.8	88.9	90.6
InfoGCN [7]	92.0	91.8	93.0	97.1	89.8	91.2
PoseC3D [9]	94.3	-	94.1	97.1	86.9	90.3
FR-Head [42]	-	-	92.8	96.8	89.5	90.9
WDCE-Net [3]	93.9	-	93.0	97.2	-	-
MDR-GCN [21]	94.5	94.2	92.8	97.2	89.8	91.3
HD-GCN [14]	94.6	94.5	93.4	97.2	90.1	91.6
DS-GCN [36]	93.4*	94.2*	93.1	97.5	89.2	91.1
BlockGCN [43]	94.5*	94.8*	93.1	97.0	90.3	91.5
HiOD (4-ensemble)	96.1	95.6	93.6	97.4	90.2	91.6
HiOD (6-ensemble)	**96.3**	**95.9**	**93.8**	**97.6**	**90.4**	**91.9**

Table 1. Comparisons of the top-1 accuracy(%) against SOTA methods on the FineGYM, FSD-10, NTU RGB+D, and NTU RGB+D 120. We use * to denote results reproduced by ourselves.

	CTR-GCN [5]	FR-Head [42]	HD-GCN [14]	Ours
NTU60-Hard	63.2	65.1	66.3	**70.2**$^{\uparrow 3.9}$
NTU60-Medium	83.8	85.4	85.1	**86.2**$^{\uparrow 1.1}$
NTU60-Easy	94.8	95.3	95.2	**95.5**$^{\uparrow 0.3}$
Reading	65.2	63.0	63.7	**68.5**$^{\uparrow 4.8}$
Writing	58.8	59.2	62.1	**66.2**$^{\uparrow 4.1}$
Play with phone	65.5	65.8	70.2	**76.0**$^{\uparrow 5.8}$
Sneeze/Cough	76.7	73.2	76.5	**80.5**$^{\uparrow 4.0}$

Table 2. Comparison of classification accuracy of three sub-datasets and four confusing action classes

Methods	S	T	Shared	Acc(%)
Baseline	✗	✗	✗	92.4
+ S	✓	✗	✗	93.0
+ T	✗	✓	✗	92.9
+ S + T	✓	✓	✗	94.1
+ Channel Split	✗	✗	✗	92.7
+ FC heads	✗	✗	✗	92.9
+ DWT	✗	✗	✗	92.8
+ S + T + Shared	✓	✓	✓	**94.4**

Table 3. Comparison of classification accuracy for effectiveness of granularity disentanglement. 'S' and 'T' represent the granularity-aware bases of spatial and temporal dimension respectively, and 'Shared' represents the shared orthogonal bases for cross-granularity interaction.

and (4) combining both spatial-temporal granularity-aware bases and shared orthogonal bases. Both the fused coarse-grained and fine-grained features are optimized using the hierarchical prototype contrastive learning strategy proposed in Section 3.3. The experimental results in Table 3 show that all these components help improve the baseline performance. Furthermore, when combined, they yield even better results, enhancing baseline performance by 2.0%.

In addition, we compare our granularity disentanglement module with three other typical disentanglement methods. 'Channel split' refers to decoupling features by dividing the channels. 'FC heads' denotes the decoupling of features by mapping the original data into different subspaces via fully connected layers. 'Wavelet' involves decomposing features into high and low frequencies using Discrete Wavelet Transform. The experimental results in Table 3 show that our method has certain advantages over the other three disentanglement methods.

Effectiveness of hierarchical prototype contrastive learning. For the optimization strategies of coarse-grained and fine-grained features, we set up four comparison schemes, as shown in Table 4. When both coarse-grained and fine-grained features are optimized using cross-entropy loss (ID 1) or hard prototype contrastive loss (ID 2), the performance is poor. Because these two loss functions are inherently 'hard' losses and cannot effectively capture the hierarchical relationships between classes. The performance improves when soft prototypes are constructed for coarse-grained features (ID 3), with optimal results achieved when the soft prototype is combined with weighted contrastive loss (ID 4). This indicates that the soft prototype provides a

ID	Methods		Acc(%)
	Coarse	Fine	
1	CE	CE	93.3
2	HP + PCL	HP + PCL	93.6
3	SP + PCL	HP + PCL	93.8
4	SP + WPCL	HP + PCL	**94.4**

Table 4. Comparison of classification accuracy for effectiveness of hierarchical prototype contrastive learning. 'CE' is cross entropy loss. 'HP' is hard prototype. 'SP' is soft prototype. 'PCL' is standard prototype contrast loss. 'WPCL' is soft prototype weighted contrastive loss.

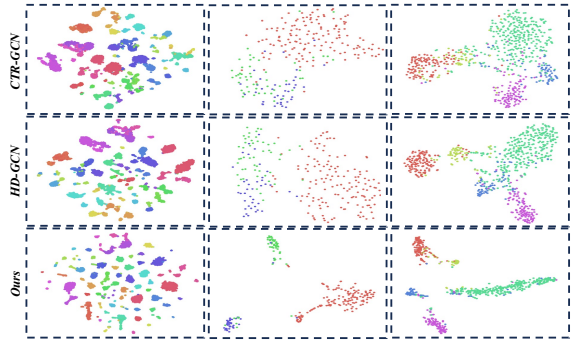

Figure 4. t-SNE visualization of CTR-GCN [5], HD-GCN [14], and our method on the FineGYM dataset. The left column shows all 99 classes, the middle column shows on classes 2, 3, and 4, while the right column shows classes 74, 75, 76, 86, and 88.

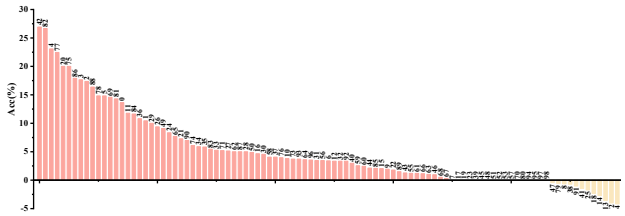

Figure 5. Accuracy difference (%) between our method and CTR-GCN [5] on 99 classes in the FineGYM dataset.

high-level semantic information to some extent, facilitating the learning of coarse-grained features. Furthermore, the weighted contrastive loss fully accounts for all similarity class prototypes, allowing the network to learn high-level hierarchical representations from samples of all classes.

4.5. Qualitative Results

t-SNE visualization. We use t-SNE to visualize some confusing action classes in the feature space, as shown in Figure 4. Compared to the other two methods [5, 14], our method more effectively delineates classification boundaries, improves intra-class compactness, and enhances the classification accuracy for fine-grained actions. In addition, Figure 5 illustrates the classification accuracy difference between our

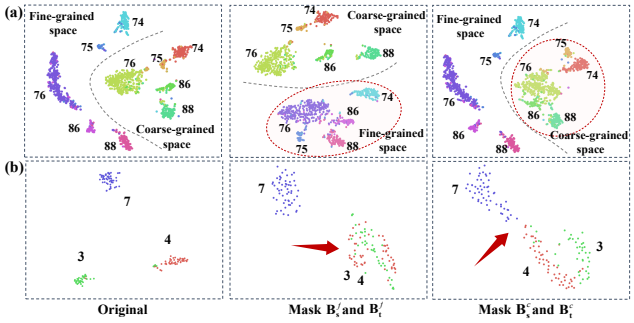

Figure 6. (a) Visualization of granularity disentanglement. (b) The impact of basis vector masking on fused features. Class 3 and Class 4 are fine-grained classes that are easily confused.

method and CTR-GCN on the FineGYM dataset, highlighting that our method significantly enhances recognition accuracy across most classes. For a detailed description of the action labels, please refer to Supplementary Material A.2.

Visualization of the role of granularity-aware bases. We visualize the disentanglement of coarse-grained and fine-grained features in Figure 6(a), and the results indicate that the coarse-grained and fine-grained features can be effectively separated. To further investigate the specific roles of different granularity-aware bases, we set 85% of the projection coefficients of the spatial and temporal coarse- or fine-grained bases to zero during the inference phase. Figure 6(a) shows that masking a granularity basis vector increases confusion in the corresponding granularity features, while the other granularity features remain unaffected. Figure 6(b) illustrates the changes in the distribution of fused features used for classification results. Notably, masking fine-grained bases makes similar classes harder to distinguish (e.g., class 3/4), while masking coarse-grained bases reduces separability between previously distinguishable classes (e.g., class 7/4). This indicates that different granularity-aware bases encode semantic information at varying levels, with coarse-grained bases capturing significant action differences and fine-grained bases focusing on subtle distinctions.

5. Conclusions

In this paper, we proposed a Hierarchical-aware Orthogonal Disentanglement (HiOD) framework for fine-grained skeleton-based action recognition. Our method disentangles coarse- and fine-grained features using independent granularity-aware bases and enhances their interaction through a cross-granularity mechanism. Additionally, hierarchical prototype contrastive learning improves coarse-grained feature learning using class hierarchy, while ensuring fine-grained distinguishability. Experiments on FineGYM, FSD-10, NTU RGB+D, and NTU RGB+D 120 datasets demonstrate the superiority of our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62332019, 62406039), the National Key Research and Development Program of China (Grant No.2023YFF1203900, 2023YFF1203903), the Beijing Nova Program (Grant No.20240484513), and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number (2023TQ0039, 2024M750257, GZC20230320).

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 1

[2] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": Fine-grained, or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11476–11485, 2021. 2, 4

[3] Haochen Chang, Jing Chen, Yilin Li, Jixiang Chen, and Xiaofeng Zhang. Wavelet-decoupling contrastive enhancement network for fine-grained skeleton-based action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4060–4064. IEEE, 2024. 1, 2, 7

[4] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4334–4342, 2021. 1

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13359–13368, 2021. 1, 2, 6, 7, 8

[6] Hao Cheng, Yufei Wang, Haoliang Li, Alex C Kot, and Bihan Wen. Disentangled feature representation for few-shot image classification. *IEEE transactions on neural networks and learning systems*, 2023. 2, 4

[7] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 1, 7

[8] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, 2022. 6

[9] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022. 7

[10] Kaixin Fan, Pengfei Ren, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, and Jianxin Liao. Pose-guided temporal enhancement for robust low-resolution hand reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22627–22637, 2025. 1

[11] Pei Geng, Xuequan Lu, Chunyu Hu, Hong Liu, and Lei Lyu. Focusing fine-grained action by self-attention-enhanced graph neural networks with contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2

[12] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2021. 2

[13] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1

[14] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2208.10741*, 2022. 1, 2, 6, 7, 8

[15] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 1, 2

[16] Ruixue Lian, William A Sethares, and Junjie Hu. Learning label hierarchy with supervised contrastive learning. *arXiv preprint arXiv:2402.00232*, 2024. 2

[17] Guangtian Liu, Haonan Su, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, and Jianxin Liao. Towards bare-hand interaction for whiteboard collaboration in virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–31, 2025. 1

[18] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 6

[19] Kaiyuan Liu, Yunheng Li, Yuanfeng Xu, Shuai Liu, and Shenglan Liu. Spatial focus attention for fine-grained skeleton-based action tasks. *IEEE Signal Processing Letters*, 29:1883–1887, 2022. 1

[20] Shenglan Liu, Xiang Liu, Gao Huang, Hong Qiao, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Ge Guo. Fsd-10: A fine-grained classification dataset for figure skating. *Neurocomputing*, 413:360–367, 2020. 6

[21] Sheng-Lan Liu, Yu-Ning Ding, Jin-Rong Zhang, Kai-Yuan Liu, Si-Fan Zhang, Fei-Long Wang, and Gao Huang. Multidimensional refinement graph convolutional network with robust decouple loss for fine-grained skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1, 2, 6, 7

[22] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 1, 7

[23] Zhendong Liu, Haifeng Xia, Tong Guo, Libo Sun, Ming Shao, and Siyu Xia. Cross-block fine-grained semantic cascade for skeleton-based sports action recognition. *arXiv preprint arXiv:2404.19383*, 2024. 1, 2

[24] Qucheng Peng, Ce Zheng, and Chen Chen. A dual-augmentor framework for domain generalization in 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2240–2249, 2024. 1

[25] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. Pose-guided hierarchical graph reasoning for 3-d hand pose estimation from a single depth image. *IEEE Transactions on Cybernetics*, 53(1):315–328, 2021. 1

[26] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. A dual-branch self-boosting framework for self-supervised 3d hand pose estimation. *IEEE Transactions on Image Processing*, 31:5052–5066, 2022. 1

[27] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20555–20565, 2022. 1

[28] Pengfei Ren, Yuchen Chen, Jiachang Hao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2163–2171, 2023. 1

[29] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8014–8025, 2023. 2

[30] Pengfei Ren, Yuanyuan Gao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Dynamic support information mining for category-agnostic pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2024. 1

[31] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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*, pages 1010–1019, 2016. 1, 2, 6

[32] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 6

[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing 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*, pages 12026–12035, 2019. 1, 7

[34] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE Transactions on image processing*, 27(7):3459–3471, 2018. 1, 2

[35] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1474–1488, 2022. 1

[36] Jianyang Xie, Yanda Meng, Yitian Zhao, Anh Nguyen, Xiaoyun Yang, and Yalin Zheng. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6225–6233, 2024. 7

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 1

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2, 7

[39] Di Yang, Yaohui Wang, Quan Kong, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Self-supervised video representation learning via latent time navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3118–3126, 2023. 2

[40] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. View-invariant skeleton action representation learning via motion retargeting. *International Journal of Computer Vision*, 132 (7):2351–2366, 2024. 2

[41] Yiwen Zhan, Yuchen Chen, Pengfei Ren, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Spatial temporal enhanced contrastive and pretext learning for skeleton-based action representation. In *Asian conference on machine learning*, pages 534–547. PMLR, 2021. 1

[42] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10608–10617, 2023. 1, 2, 7

[43] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024. 7

[44] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 1, 2

[45] Zixin Zhu, Le Wang, Wei Tang, Ziyi Liu, Nanning Zheng, and Gang Hua. Learning disentangled classification and localization representations for temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3644–3652, 2022. 2