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

Supplementary Material

In the supplemental material, we provide:

- more implementation details in Sec. A,
- more qualitative results in Sec. B,
- more quantitative results in Sec. C.

A. Implementation Details

A.1. Details of K-means hierarchical discovery.

Since most action recognition datasets only provide single-level labels, we consider using deep clustering algorithm to discover the hierarchical structure of classes within the dataset. Specifically, we use the traditional K-means algorithm to divide all samples into K clusters. The clustering assignment matrix $\mathbf{S} \in \mathbb{R}^{K \times N_c}$ represents the hierarchical relationship of classes, where the element $\mathbf{S}_{k,n}$ indicates the number of samples from the n-th class in the k-th parent cluster.

A.2. Training Details

Backbone. In this paper, we use HD-GCN [4] as our backbone. Specifically, we employ a 10-layer network, where each layer consists of an HD-Gconv block and a multi-scale temporal convolution block. The output channel configuration for each layer is set to 64-64-64-64-128-128-128-256-256-256. Theoretically, our proposed Hierarchical-aware Orthogonal Disentanglement framework (HiOD) is applicable to any skeleton-based action recognition backbone. In Section C.5, we also present experimental results using different backbones.

FineGYM dataset details. We use the method provided in Pyskl [3] to obtain the 2D skeleton sequences from the FineGYM dataset [7]. The skeleton contains 17 joints, with 128 input frames and 3 input channels. The third dimension of the channel represents the confidence score. The accuracy of various methods on the FineGYM dataset is primarily sourced from previous works [1, 5], while the accuracy of the HD-GCN method is based on our experimental results. In our experiments, we report the Top-1 accuracy of our method on the FineGYM dataset.

FineGYM class label. We present the classes of gymnastic actions and their corresponding label descriptions involved in the t-SNE visualization experiment in Table 1. Classes 74, 75 and 76 are fine-grained action classes that are easily confused, and classes 2, 3 and 4 are also fine-grained action classes prone to confusion.

FSD-10 dataset details. The skeleton of FSD-10 dataset contains 25 joints, with 1500 input frames and 3 input channels. We follow the data preprocessing method proposed in [5]. To balance speed and accuracy, we set the size of the random cropping window to 600.

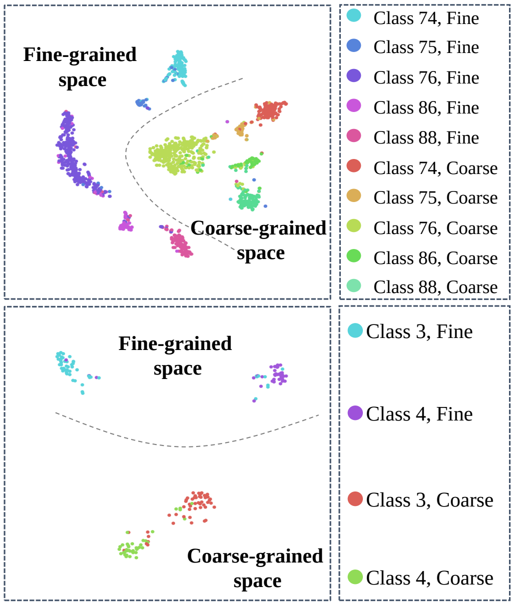

Figure 1. t-SNE visualization of granularity disentanglement. 'Fine' represents the disentangled fine-grained features, while 'Coarse' represents the disentangled coarse-grained features. Detailed descriptions of all class labels can be found in Table 1.

B. More Qualitative Results

B.1. Granularity Disentanglement Results

To demonstrate the effectiveness of our method in achieving granularity disentanglement, we utilize t-SNE to visualize the distribution of coarse-grained and fine-grained features in the high-level semantic space on the FineGYM [7] dataset, as illustrated in Figure 1. The results indicate that our method effectively maintains a clear separation between the coarse-grained and fine-grained feature spaces, thereby mitigating the interference of coarse-grained features with fine-grained features during training. Furthermore, we observe that fine-grained features of easily confused actions exhibit superior inter-class separability compared to coarse-grained features, a characteristic that is essential for fine-grained action recognition. This benefits from the hierarchical prototype contrastive learning strategy. Specifically, the hard prototype contrastive loss ensures the ac-

Class id	Event	Label
2	Vaulting Horse	round-off, flic-flac on, stretched salto backward **with 1 turn off**
3	Vaulting Horse	round-off, flic-flac on, stretched salto backward **with 1.5 turn off**
4	Vaulting Horse	round-off, flic-flac on, stretched salto backward **with 2.5 turn off**
7	Floor Exercise	switch leap with 1 turn
74	Uneven Bar	pike sole circle backward **with 1 turn to handstand**
75	Uneven Bar	pike sole circle backward **with 0.5 turn to handstand**
76	Uneven Bar	pike sole circle backward **to handstand**
86	Uneven Bar	clear pike circle backward to handstand
88	Uneven Bar	stalder backward to handstand

Table 1. Description of the classes and labels involved in t-SNE visualization. Classes 74, 75 and 76 are fine-grained action classes that are easily confused, and classes 2, 3 and 4 are also fine-grained action classes prone to confusion.

curacy of fine-grained prototypes, promoting the separation of fine-grained features. Meanwhile, the soft prototype weighted contrastive loss enables coarse-grained features to learn shared semantic representations from similar classes, resulting in inter-class similarity. Consequently, coarse-grained features effectively represent higher-level parent classes, which in turn implicitly facilitates the learning of fine-grained features.

B.2. Class Activation Map Visualization

We employed Grad-CAM [6] to visualize the coarse-grained and fine-grained features, as shown in Figure 2. CAM highlights the class-specific discriminative regions, i.e. specific joints and frames. The results show that coarse-grained features tend to focus on a larger range of frame sequences and more joints, while fine-grained features concentrate on more specific areas. This suggests that coarse-grained features capture the global motion patterns of high-level parent classes, while fine-grained features focus on local variations in detail. Compared with the HD-GCN, our method can better locate discriminative regions, both for coarse-grained and fine-grained features. In addition, we observed that as classification difficulty increases (from left to right in Figure 2), the distinction between coarse and fine-grained features becomes more pronounced. This demonstrates that our granularity disentanglement strategy effectively discovers the discriminative details and emphasizes their significance.

C. More Quantitative Experiments

C.1. Effect of basis vector number

We study the effect of changing the number of basis vectors in the orthogonal basis on action recognition. The experimental results are detailed in Table 2. We set the number of basis vectors to 20, 40, 60, 80, 100, and 120. The experimental results show that the best effect is achieved when the number is 80. This may be attributed to the weaker feature representation when the number of basis vectors is low, and the tendency for the orthogonal basis to capture redundant information when the count is high.

Number of basis vectors	Acc(%)
20	93.7
40	93.9
60	94.1
80	94.2
100	**94.4**
120	94.0
140	93.6

Table 2. Comparison of classification accuracy for number of basis vectors.

Methods	Acc(%)
w/o refine	93.8
MLP refine (1 layer)	93.9
MLP refine (2 layers)	94.0
MLP refine (4 layers)	94.2
MLP refine (6 layers)	94.1
GCN-TCN refine (1 layer)	94.2
GCN-TCN refine (2 layers)	**94.4**
GCN-TCN refine (3 layers)	94.3

Table 3. Comparison of classification accuracy for effectiveness of refine module.

C.2. Effectiveness of Refine Module

To investigate the specific role of the feature refinement module, we conducted ablation experiments on the configuration of the refinement module, with the results shown in Table 3. We constructed our refinement module using either MLPs or GCN-TCN layers. The MLP refines the classification vector obtained from global average pooling, while the GCN-TCN layer refines the fused features $\mathbf{X}_{fuse}$. We

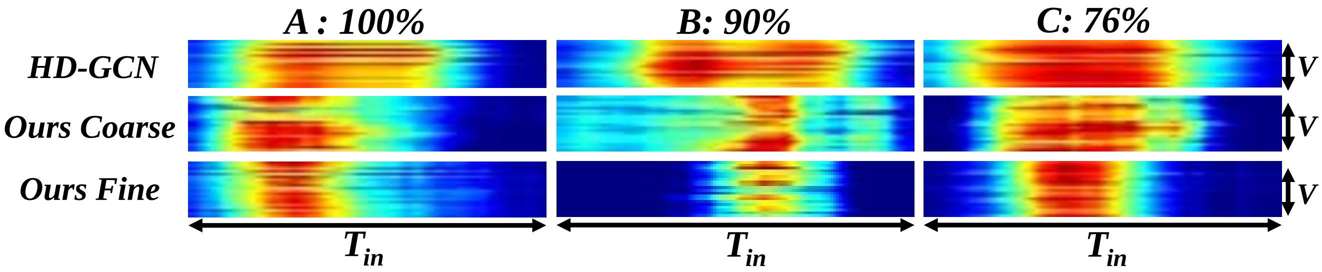

Figure 2. From top to bottom, they are CAMs for HD-GCN [4], our coarse-grained features, and our fine-grained features. From left to right, they are three action samples organized by recognition difficulty, from easy to difficult, along with the recognition accuracy for each class.

Methods	B^c_s	B^c_t	B^f_s	B^f_t	B^{shared}	Acc(%)
Spatial-coarse	✓				✓	91.8
Spatial-fine			✓		✓	93.4
Temporal-coarse		✓			✓	86.7
Temporal-fine				✓	✓	77.9
Spatial	✓		✓		✓	93.5
Temporal		✓		✓	✓	91.4
Without shared bases	✓	✓	✓	✓		92.3
All	✓	✓	✓	✓	✓	**94.4**

Table 4. Comparison of top-1 accuracy using different orthogonal bases during the inference phase on the FineGYM dataset.

observed that the performance was the worst without the refinement module, and the type of refinement module had little impact on accuracy. This indicates that the refinement module primarily serves to stabilize the feature decoupling process and enhance the independence of multi-granularity features, rather than relying on a specific network architecture.

C.3. Importance of Different Orthogonal Bases

During the inference phase, we quantitatively investigate the importance of each orthogonal basis for action recognition using the basis vector masking strategy mentioned in Section 4.5 of the main text. Specifically, we mask 100% of the projection coefficients of a particular orthogonal basis to eliminate its contribution to action recognition. Table 4 presents the specific experimental results. The findings indicate that spatial granularity-aware bases contribute more significantly to fine-grained action recognition compared to temporal granularity-aware bases. In the spatial dimension, fine-grained bases are more important than coarse-grained bases, while the opposite is true in the temporal dimension. The removal of shared bases results in a decrease in recognition accuracy, which further illustrates the importance of shared bases.

Methods	Acc (%)
Hierarchical prototypes	93.8
Soft prototypes	**94.4**

Table 5. Comparison of top-1 accuracy using soft prototypes and hierarchical prototypes on the FineGYM dataset.

C.4. Comparison between Soft Prototypes and Hierarchical Prototypes

In the hierarchical contrastive learning phase, we compared the differences between soft prototypes and hierarchical prototypes. The proposed soft prototype maintains a prototype for each action class using coarse-grained features, with the number of soft prototypes equal to the total number of classes. In contrast, hierarchical prototypes maintain a prototype for each parent class, where parent classes are obtained through K-means clustering, and the number of hierarchical prototypes equals the number of clusters K. Table 5 shows that using soft prototypes in hierarchical contrastive learning can achieve higher classification accuracy. This is because soft prototypes not only capture shared parent-class features among similar classes but also preserve certain subclass differences, resulting in more powerful feature representation capabilities.

C.5. Comparison Experiments Using Different Backbones

To evaluate the generality and robustness of the proposed Hierarchical-aware Orthogonal Disentanglement framework (HiOD), we conducted comparative experiments using various backbone networks. The experimental results are shown in Table 6. The results indicate that our HiOD method improves accuracy across all backbones, demonstrating the generalization capability of our approach.

References

[1] Haochen Chang, Jing Chen, Yilin Li, Jixiang Chen, and Xiaofeng Zhang. Wavelet-decoupling contrastive enhancement

Backbone	Acc(%)
ST-GCN [8]	87.1
+ HiOD	91.2
CTR-GCN [2]	90.5
+ HiOD	92.1
WDCE-Net [1]	90.4
+ HiOD	92.4
HD-GCN [4]	92.4
+ HiOD	**94.4**

Table 6. Comparison of top-1 accuracy using different backbones on the FineGYM dataset.

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, 4

[2] 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. 4

[3] 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. 1

[4] 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, 3, 4

[5] 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

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[7] 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. 1

[8] 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. 4