(APPENDIX) BReSK: Bootstrapped Contrastive Representation Learning for Skeleton-Based Action Understanding

Anonymous CVPR submission

Paper ID 12

Appendix Overview

The Supplementary material is organized as follows:

- Section 1: Datasets and Evaluation Metrics
- Section 2: Evaluation Protocols
- Section 3: Framework and Implementation Details
- Section 4: Additional SOTA Comparisons
  1. Action recognition (Multi-Modality) -Linear Evaluation
  2. Action Retrieval - KNN Evaluation
  3. Action Prediction Evaluation
- Section 5: Additional Ablation Studies
  1. Disentanglement process
  2. Deeper Insights, Analysis, and Comparison
  3. Representation visualization using t-SNE

1. Datasets and Evaluation Metrics

NTU-RGB+D 60 [17] consists of 56,880 sequences of high-quality 3D skeletons, captured by the Microsoft Kinect v2 sensor. We only use sequences of 3D skeletons in this work, and we follow the cross-subject (CS) and cross-view (CV) evaluation protocol.

NTU-RGB+D 120 [13] is an extension of NTU-RGB+D 60, increasing the number of action classes from 60 to 120. The dataset contains a total of 114,480 video samples. We use 3D skeleton sequences and follow the Cross-Subject (CS) and Cross-Setup (CSet) evaluation protocols.

PKU-MMD [3] : For 3D skeleton action classification on PKU-MMD, action instances are segmented based on temporal annotations and are partitioned into training and testing sets according to the cross-subject evaluation protocol. In our experiments, we use PKU-MMD II, with 5,332 samples for training and 1,613 samples for testing.

Posetics [28] is built upon Kinetics400 videos, the dataset consists of 142,000 real-world video clips across 320 action classes, along with corresponding 2D and 3D skeletons. We leverage the Posetics dataset to pre-train our action representation learning framework using skeleton data and explore transfer learning for skeleton-based action classification. Top-1 and Top-5 accuracy are used as evaluation metrics.

Toyota Smarthome [4] (SmartHome) is a real-world collection for daily living action classification, containing 16,115 videos across 31 action classes. It includes RGB videos, as well as 2D and 3D skeleton data [27]. Since the 2D data is more robust for action recognition, even in cross-view evaluations, we use 2D data for the experiments unless otherwise specified. For evaluation, we report the mean per-class accuracy following the cross-subject (CS) and cross-view (CV2) protocols.

Penn Action [32] contains 2,326 video sequences of 15 different real-world sports actions. In this work, we use 2D skeletons obtained by LCRNet++ for our experiments, and we report Top-1 accuracy following the standard train-test splitting.

2. Evaluation Protocols

We evaluate BReSK under four complementary self-supervised learning protocols following prior works [1, 12, 24, 25, 29, 31]. (i) We conduct **linear evaluation** to measure the intrinsic discriminability of the learned representations using a frozen encoder. (ii) We perform **cross-dataset transfer** by pretraining on a source dataset and adapting to unseen target datasets to assess generalization. (iii)We study **semi-supervised learning** by fine-tuning with only 5% or 10% labeled samples to evaluate data efficiency. (iv) We evaluate temporal reasoning through **action prediction** from partial observations (70%, 80%, 90% of the sequence). (iv) We examine **action detection** in untrimmed sequences using mAP at multiple temporal IoU thresholds (0.1, 0.3, 0.5). Together, these protocols assess representation discriminability,

Category	Parameter	Lab setting & Real-world
Optimizer	Optimizer type	SGD
	Momentum	0.9
DiP Settings	num-layers	1
	Prediction-dim	128
Contrastive Settings	Queue size	8192
	Momentum	0.999
	Softmax temperature	0.2
	Projection dimension	128
Training Setup	Weight decay	0.0001
	Initial learning rate	0.01
	Learning rate schedule	[350]
	Batch size	64
	λ	[100 - 1]
	Training steps	440

Table 1. Training settings for self-supervised learning on the lab setting and real-world (NTU-RGB+D and Posetics dataset).

Parameter	Action Classification		Action Retrieval & Action Prediction	Transfer Learning		Semi-Supervised		Action Detection
	Lab	Real		Lab	Real	Lab	Real	Lab
Initial learning rate	2		2	0.2	2	0.2		0.01
Batch size	1024		1024	32	1024	32	16	32
Training steps	80		80	80		80	100	60

Table 2. Evaluation hyperparameter across different tasks and settings (lab and real-world) for various datasets.

data efficiency, cross-domain generalization, early action understanding, and temporal localization ability.

3. Framework and Implementation Details

Following prior state-of-the-art works [5, 25], all input skeleton sequences are pre-processed to a fixed length of 64 frames. We then apply the same data augmentation strategy as in [25] to generate two different views of the input, serving as the query and key sequences.

The query and key branches share the same architecture and functionality following MoCo-v2 [6]. Below, we describe the query branch in detail.

We adopt a lightweight version of CTR-GCN [2] composed of 3 blocks as our backbone for skeleton feature extraction, with the output channel dimension fixed at 64. The extracted features are then separated into motion view (fm) and joints view (fj) components after reshaping and are projected into a higher semantic space.

These features are followed by Dual-Branch Instance Predictor **DiP**, which is an implicit lightweight feature refinement module designed to enhance representation quality through cross-view interaction. It adopts an asymmetric dual-branch architecture to enforce cross-view consistency while avoiding representation collapse.

The query branch is equipped with lightweight predictor modules, namely **Pred-S** and **Pred-T**, to model cross-view relationships between spatial and temporal representations. Let $q_i \in \mathbb{R}^d$ denote the query representation from view $i \in \{s, t\}$, where s and t correspond to spatial and temporal streams, respectively. The predictor $f_{i\to j}(\cdot)$ maps features from view i to the feature space of view j. Each predictor is implemented as a two-layer MLP.

Finally, Bootstrapped Contrastive Learning (BoCL) loss is applied to enhance the representation by jointly optimizing two complementary objectives, i.e., a *Cross-Branch Prediction Alignment Loss* ($\mathcal{L}_{\text{CroP}}$) that enforces cross-view semantic consistency, and a *Mixed Contrastive Loss* ($\mathcal{L}_{\text{MiCo}}$) that sharpens inter-instance discrimination via a momentum-updated memory bank.

The MoCo parameters are $m = 0.999$ for the EMA update parameter, and $k = 8192$ for the queue length. We pretrain our model with SGD optimizer for 440 epochs with a batch size of 64, and we pretrain our framework on one Nvidia H100 GPU. For the evaluation and downstream task, we use the trained query branch encoder under different protocol evaluations. All hyperparameters used during training and evaluation are detailed in Tables 1 and 2, respectively.

4. Additional SOTA Comparisons

In this section, we provide additional state-of-the-art (SOTA) comparisons and evaluations across both real-world and lab-setting (controlled) datasets. For the lab-setting datasets, we

Method	NTU-RGB+D 60 CS(%)		NTU-RGB+D 60 CV(%)		NTU-RGB+D 120 CS(%)		NTU-RGB+D 120 CSet(%)	
	J	J+B+M	J	J+B+M	J	J+B+M	J	J+B+M
CrosSCLR [10] (CVPR 21)	72.9	77.8	79.9	83.4	-	67.9	-	66.7
CMD [14] (ECCV 22)	79.8	84.1	86.9	90.9	70.3	74.7	71.5	76.1
ActCLR [11] (CVPR 23)	80.9	84.3	86.7	88.8	69.0	74.3	70.5	75.7
AimCLR [20] (AAAI 22)	74.3	78.9	79.7	83.8	63.4	68.2	63.4	68.8
RVTCLR+ [33] (ICCV 23)	74.7	79.7	79.1	84.6	-	60.0	-	68.9
UmURL [18] (ACMMM 23)	82.3	84.2	89.8	90.0	73.5	75.2	74.3	76.3
PCM^3 [30] (ACMMM 23)	83.9	87.4	90.4	93.1	76.5	80.0	77.5	81.2
HiCo [5] (AAAI 23)	81.1	83.8	88.6	90.8	72.8	75.9	74.1	77.3
SCD-Net [25] (AAAI 24)	86.6	87.3	91.7	-	76.9	-	80.1	-
USDRL [24] (AAAI 25)	84.2	87.1	90.8	93.2	76.0	79.3	76.9	80.6
BReSK (Our)	**87.3**	**88.2**	**92.0**	**93.2**	**79.1**	**81.8**	**80.1**	**82.5**

Table 3. **Action Recognition (Multi-Modality)**- *Linear Evaluation*- for Lab-setting datasets (NTU-RGB+D 60 and NTU-RGB+D 120)

Methods	Action Retrieval	
	Posetics Top-1 (%)	PennAction Top-1 (%))
HiCo [5](AAAI 23)	9.6	70.8
PCM^3 [30] (ACMMM 23)	7.6	73.7
ViA [29] (IJCV 24)	-	76.9
USDRL [24](AAAI 25)	12.7	76.7
BReSK-2D (Our)	**13.3**	**78.4**
HiCo [5](AAAI 23)	15.9	86.8
USDRL [24](AAAI 25)	20.2	89.8
BReSK-3D (Our)	**20.8**	**90.2**

Table 4. **Action Recognition 2D and 3D skeleton** – Combined results for *Linear Evaluation*, *Action Retrieval* and *Transfer Learning* on Real-world datasets (Posetics, SmartHome, and PennAction).

Method	70%	80%	90%	100%
DeepSCN [9]	58.2	60.2	60.0	58.6
MSRNN [7]	63.9	67.4	68.9	69.2
P-TSL [23]	77.6	80.1	81.5	82.0
SCDNet [25]	79.5	82.1	85.0	86.6
USDRL (STTR) [24]	80.3	82.4	83.7	84.2
USDRL (DSTE) [24]	81.4	83.6	84.5	85.2
BReSK-3D (Ours)	**82.0**	**84.2**	**86.1**	**87.3**

Table 5. **Action Prediction**- *Linear Evaluation results*- for Lab-setting datasets (NTU-RGB+D60 CS).

Method	70%	80%	90%	100%
Penn Action				
HiCo [5]	77.8	80.6	83.4	87.6
PCM^3 [30]	53.4	56.7	62.4	85.7
USDRL [24]	85.7	86.9	87.9	89.6
BReSK-2D (Our)	**86.7**	**89.2**	**90.8**	**91.9**
Posetics				
HiCo [5]	18.5	19.9	20.5	21.3
PCM^3 [30]	17.1	18.1	18.7	20.0
USDRL [24]	24.2	24.8	25.4	25.9
BReSK-2D (Our)	**27.6**	**28.5**	**29.2**	**29.8**

Table 6. **Action Prediction**- *Linear Evaluation results*- on Real-world datasets (Penn Action and Posetics).

Method	Transfer to PKU-MMD II
	NTU-RGB+D 120 CS(%)
HiCo-Transformer [5]	55.4
CrosSCLR-B [10]	52.8
CMD [14]	57.0
UmURL-3 [18]	58.5
A^{2}MC [26]	58.9
USDRL+ [22]	58.3
Heterogeneous [21]	63.1
BReSK-3D (Our)	**64.2**

Table 7. **Action Recognition**- *Transfer Learning Evaluation (Fine-tune) results*- for Lab-setting datasets (from NTU-RGB+D 120 to PKU-MMD II).

present additional experiments, including multimodal linear evaluation and action retrieval. For real-world scenarios, we adopt recent SOTA methods [5, 24] on 2D and 3D skeleton sequences to ensure a fair comparison with our approach. We evaluate action retrieval after being pretrained on Posetics dataset.

4.1. Action recognition (Multi-Modality) -Linear Evaluation:

Skeleton sequence can be represented in different modalities: joint-based (J), focusing on individual joint positions to capture static postures; bone-based (B), encoding the relative positions between connected joints to highlight body part interactions; and motion-based (M), capturing the changes in joint or bone positions over time to emphasize dynamic movement patterns. These modalities offer complementary insights into both static and dynamic aspects of human movement. In Table 3, we report the linear evaluation results compared to SoTA on lab-setting datasets (NTU-RGB+D 60 and NTU-RGB+D 120) using joint-based and multi-modality (J+B+M). For the multi-modality case, we first train our model separately for each modality and then fuse the results from all three modalities. BReSK shows a significant improvement compared to the recent SoTA [24, 25, 30], across

Method	NTU-RGB+D 60		NTU-RGB+D 120	
	CS (%)	CV (%)	CS (%)	Cset (%)
ISC [19]	62.5	82.6	50.6	53.0
CrosSCLR-B [10]	66.1	81.3	52.2	54.9
HaLP [16]	65.8	83.6	55.8	59.0
SkeAttnCLR [8]	69.4	67.8	46.7	58.0
HiCo [5]	68.3	84.8	56.6	59.1
MAMP [15]	62.0	70.0	51.8	56.1
A^{2}MC [26]	70.8	85.4	59.1	62.6
Heterogeneous [21]	66.3	87.1	55.7	59.8
BReSK-3D (Our)	**75.2**	**87.2**	**61.8**	**64.8**

Table 8. **Action retrieval**- *KNN Evaluation results*- for Lab-setting datasets (NTU-RGB+D 60 & NTU-RGB+D 120).

Activity	Gain from BReSK-2D
Usetablet	+64.67
Walk	+29.23
Cutbread	+21.82
Eat.Snack	+19.43
Drink.Fromcan	+17.05
Pour.Frombottle	+10.83
Uselaptop	+9.54
Drink.Fromcup	-1.05
Sitdown	-2.00
Drink.Frombottle	-3.95
Pour.Fromcan	-6.66
Mean Accuracy	**+4.3**

Table 9. Gain from BReSK compared to ViA on SmartHome-CV2 dataset. The results are Mean per class.

all benchmark datasets with different types of modality.

4.2. Action Prediction - Linear Evaluation

Tables 5 and 6 report BReSK's performance on action prediction, demonstrating its ability to anticipate actions from partial observations. Thanks to the design of BReSK, which allows for capturing robust spatio-temporal dependencies, enabling accurate early recognition even when only a fraction of the skeleton sequence is available. This highlights the model's strength in temporal reasoning and its potential for real-time applications, where predicting ongoing actions from limited observations is critical.

4.3. Action Retrieval - KNN Evaluation

Table 4 and Table 8 report the action retrieval performance using a K-Nearest Neighbor (KNN) evaluation on (Posetics, PennAction, NTU-RGB+D 60 ,and NTU-RGB+D 120). We use a K-Nearest Neighbor (KNN) classifier, a non-parametric supervised learning method, which directly reflects the quality of the representation space learned by the frozen encoder. BReSK consistently outperforms recent state-of-the-art methods, including Heterogeneous [21] and A^{2}MC [26],

Activity	Gain from BReSK-2D
Drink.Fromcup	+50.91
Leave	+47.23
Readbook	+45.09
Pour.Fromkettle	+36.99
Uselaptop	+31.41
Enter	+26.28
Usetelephone	+22.20
Pour.Fromcan	-2.30
Drink.Fromglass	-2.30
Maketea.Boilwater	-2.30
Drink.Frombottle	-4.57
Makecoffee.Pourgrains	-6.05
Mean Accuracy	**+7.0**

Table 10. Gain from BReSK compared to ViA [29] on SmartHome-CS dataset. The results are Mean per class.

across all evaluation protocols. While prior approaches such as HiCo [5] and A^{2}MC demonstrate strong performance under specific settings, their results are less consistent across different protocols. In contrast, BReSK maintains high accuracy across all evaluations, indicating that the learned representations are both robust and transferable. These results highlight the effectiveness of BReSK in learning discriminative feature embeddings for skeleton-based action retrieval. As well as demonstrating strong generalization capability and effective cross-dataset transfer of the learned representations.

5. Additional Ablation Studies

5.1. Disentanglement process (Spatial and Temporal)

The decoupling process is considered an essential step and has proven effective in many SOTA. While USDRL [24] relies on a dense spatio-temporal encoder and mask strategy applied directly to the projected embeddings from reshaped skeleton sequences, and ViA [29] leverages orthogonal basis vectors to disentangle motion in the feature space, BReSK applies lightweight projection layers directly to intermediate features after reshaping from a motion (time) and joint perspective. This design makes the features more effective and adaptable in our framework.

Specifically, given a skeleton sequence $\mathbf{SK} \in \mathbb{R}^{C\times T\times V\times M}$, with C 2D/3D, temporal length T and V skeletal joints, intermediate features $\mathbf{f}\text{sk} \in \mathbb{R}^{C\text{inter}\times T\times V}$ are first extracted. These features are then transformed into two distinct views (motion and joint) by reshaping them into $\mathbf{f}m \in \mathbb{R}^{T\times(V\times C\text{in})}$ for the motion domain and $\mathbf{f}j \in \mathbb{R}^{V\times(T\times C\text{in})}$ for the joint domain. The rearranged sequences are subsequently mapped into a shared embedding space using separate MLPs (Eq. 1), producing feature representations $\mathbf{F}_m \in \mathbb{R}^{T\times C_e}$ and $\mathbf{F}_j \in \mathbb{R}^{V\times C_e}$. This decou-

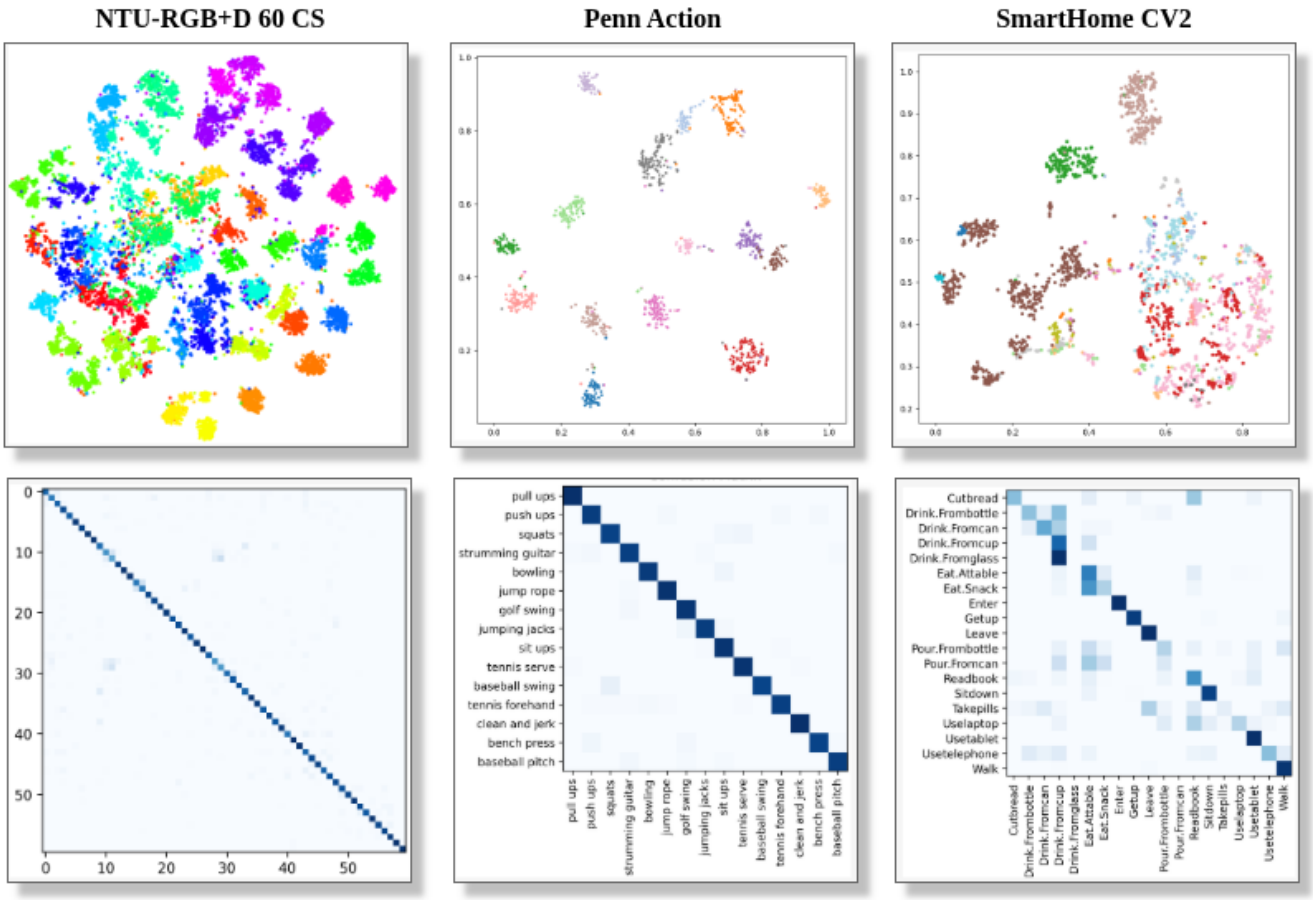

Figure 1. Confusion matrix and TSNE plotof BReSK features NTU-RGB+D 60, Penn CS datasets.

pling enables actions to be represented from complementary perspectives.

5.2. Deeper Insights, Analysis and Comparison

The class-wise evaluation across SmartHome-CS and SmartHome-CV2 provides deeper insights into the inductive biases of BReSK and its ability to generalize under cross-subject and cross-view conditions.

Table 9 shows that BReSK-2D improves over ViA with a +4.3 mean accuracy gain on the **SmartHome-CV2 dataset**, indicating overall better performance. The largest improvements are observed for dynamic and subtle actions, such as Usetable (+64.67), Walk (+29.23), and Cutbread (+21.82), as well as Eat.Snack and Drink.Fromcan. This suggests that BReSK-2D is effective at capturing fine-grained motion patterns and temporal dynamics from skeleton data.

In contrast, slight performance drops are observed for a few classes (e.g., Drink.Fromcup, Pour.Fromcan), likely due to high similarity between motion patterns and limited discriminative cues in skeleton representations.

on **SmartHome-CS** dataset, Table 10 shows that BReSK-2D achieves a larger overall improvement , with a +7.0 mean accuracy gain, confirming the consistency of its advantage over ViA. Similar to the previous results, the strongest gains appear in subtle and dynamic actions, such as Drink.Fromcup (+50.91), Leave (+47.23), Readbook (+45.09), and Pour.Fromkettle (+36.99). Activities like Uselaptop, Enter, and Usetelephone also show notable improvements, highlighting the ability of BReSK-2D to capture fine-grained motion dynamics from skeleton data.

However, performance drops are observed for several classes (e.g., Pour.Fromcan, Drink.Fromglass, Drink.Frombottle, Makecoffee.Pourgrains). These activities are object-centric and involve similar interaction patterns, where distinguishing the action depends heavily on object information. Since skeleton data does not encode object cues, these cases remain challenging.

Taken together, these results confirm that BReSK-2D consistently excels at capturing subtle and dynamic motion patterns from skeleton data, leading to strong improvements across diverse activities. At the same time, its limitations are evident in object-centric actions, where the absence of explicit object information makes it difficult to distinguish between similar interaction patterns, highlighting an inherent

CVPR #12

challenge of skeleton-based representations.

5.3. Representation visualization using t-SNE

To further analyze the quality of the representations learned by BReSK, we conduct a qualitative and quantitative evaluation across three benchmark datasets: NTU-RGB+D 60 CS, Penn Action, and SmartHome CV2. Figure 1 illustrates the t-SNE projections of the extracted features alongside their corresponding confusion matrices, providing insight into the discriminative capacity of our model under varying dataset complexities.

We first assess the structure of the learned embedding space using t-SNE projections (Figure 1, top row). On NTU-RGB+D 60 CS, which comprises 60 action categories, BReSK produces a large number of visually distinct clusters, demonstrating its capacity to maintain inter-class separability even under high class-count conditions. On Penn Action, the 15 action classes form tightly grouped and clearly separated clusters, indicating that BReSK captures highly consistent intra-class features while preserving strong inter-class margins. On the more challenging SmartHome CV2 dataset, where activities are semantically fine-grained and visually similar, partial cluster overlaps are observed — particularly among actions sharing similar objects or motion patterns. This behavior is expected and reflects the inherent difficulty of fine-grained activity recognition rather than a limitation of the learned representations.

The confusion matrices (Figure 1, bottom row) complement the t-SNE analysis by quantifying per-class prediction accuracy. Across all three datasets, a dominant diagonal structure is consistently observed, confirming that BReSK correctly classifies the majority of samples for each action class. On NTU-RGB+D 60 CS, the diagonal remains strong across all 60 classes with minimal off-diagonal activations. On Penn Action, the near-perfect diagonal corroborates the clean cluster separation seen in the embedding space. On SmartHome CV2, localized off-diagonal entries appear between semantically related classes, such as Drink.Frombottle and Drink.Fromcan, which are distinguishable only by subtle object-level differences, highlighting the challenging nature of this benchmark.

The consistency between the t-SNE visualizations and the confusion matrices across all three benchmarks confirms that BReSK learns compact, discriminative, and generalizable feature representations. The progressive increase in classification complexity from Penn Action to NTU-RGB+D 60 CS to SmartHome CV2 is clearly reflected in both the embedding structure and the confusion patterns, validating that our model scales gracefully with dataset difficulty. These results collectively support the design choices underlying BReSK and justify its effectiveness as a robust skeleton-based action recognition framework.

References

[1] Wenming Cao, Liangxi Qian, Yicha Zhang, Xuelong Li, and Xinpeng Yin. Asymmetric context-guided adaptive alignment network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[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. 2

[3] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv:1703.07475*, 2017. 1

[4] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, 2019. 1

[5] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *AAAI*, 2023. 2, 3, 4

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[7] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang. Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[8] Yilei Hua, Wenhan Wu, Ce Zheng, Aidong Lu, Mengyuan Liu, Chen Chen, and Shiqian Wu. Part aware contrastive learning for self-supervised action recognition. In *IJCAI*, 2023. 4

[9] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[10] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *(CVPR)*, 2021. 3, 4

[11] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[12] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Self-supervised skeleton representation learning via actionlet contrast and reconstruct. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[13] 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 TPAMI*, 2020. 1

[14] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *ECCV*, 2022. 3

[15] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *ICCV*, 2023. 4

[16] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Kumar Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions, 2023. 4

[17] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3D human activity analysis. *CVPR*, 2016. 1

[18] Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2973–2984, 2023. 3

[19] Fida Mohammad Thoker, Hazel Doughty, and Cees G. M. Snoek. Skeleton-contrastive 3d action representation learning, 2021. 4

[20] Guo Tianyu, Liu Hong, Chen Zhan, Liu Mengyuan, Wang Tao, and Ding Runwei. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI*, 2022. 3

[21] Hongsong Wang, Xiaoyan Ma, Jidong Kuang, and Jie Gui. Heterogeneous skeleton-based action representation learning. In *(CVPR)*, 2025. 3, 4

[22] Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo sen Xie, Xin Geng, and Liang Wang. Foundation model for skeleton-based human action understanding. *Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[23] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[24] Wanjiang Weng, Hongsong Wang, Junbo Wang, Lei He, and Guosen Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1, 3, 4

[25] Cong Wu, Xiao-Jun Wu, Josef Kittler, Tianyang Xu, Sara Atito, Muhammad Awais, and Zhenhua Feng. Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *AAAI*, 2024. 1, 2, 3

[26] Binqian Xu, Xiangbo Shu, Jiachao Zhang, Rui Yan, and Guo-Sen Xie. Attack-augmentation mixing-contrastive skeletal representation learning, 2024. 3, 4

[27] Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, and Francois Bremond. Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos. In *WACV*, 2021. 1

[28] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Unik: A unified framework for real-world skeleton-based action recognition. In *BMVC*, 2021. 1

[29] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Via: View-invariant skeleton action representation learning via motion retargeting. *IJCV*, 2024. 1, 3, 4

[30] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 3

[31] Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. Self-supervised skeleton-based action representation learning: A benchmark and beyond. *IJCV*, 2026. 1

[32] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 1

[33] Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3