CVPR #12

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

Anonymous CVPR submission

Paper ID 12

Abstract

Self-supervised learning especially contrastive learning has emerged as a powerful paradigm for skeleton-based action recognition. However, existing approaches often rely on heavy architectural refinements, such as transformer-based modules, which introduce redundancy and increase model complexity without necessarily improving representation consistency. To address this issue, we propose BReSK, a self-supervised framework that combines a bootstrap prediction with momentum contrastive learning for skeleton-based action understanding. At the core of BReSK is DiP, an asymmetric dual-branch predictor that enforces cross-view consistency through spatial and temporal predictors in the Query branch, while using exponentially moving averaged (EMA) targets in the Key branch to stabilize representation learning. In addition, we introduce BoCL, a hybrid objective that jointly optimizes a bootstrap alignment loss ($\mathcal{L}_{CroP}$) and a momentum-based contrastive loss ($\mathcal{L}_{MiCo}$), improving instance discrimination while reducing class confusion in the embedding space. Extensive experiments on six benchmark datasets, including NTU-RGB+D 60/120, PKU-MMD, Toyota SmartHome, Penn Action, and Posetics, show that BReSK consistently outperforms state-of-the-art methods across diverse settings while using fewer parameters.

1. Introduction

Human Action Understanding (HAU) [1, 3, 6, 18, 25, 26, 30, 35] is fundamental for applications such as healthcare monitoring, daily activity analysis, and sports understanding. Among various modalities, skeleton-based action recognition is particularly effective due to its robustness to appearance variations and its ability to model structured human motion [3, 6, 18, 30]. Recent advances have been largely driven by self-supervised learning (SSL), especially contrastive learning methods [16, 19, 28]. These approaches learn discriminative representations by aligning augmented views of the same skeleton while separating different samples, enabling effective learning from large-scale unlabeled data and improving generalization over supervised methods [8, 12, 16, 19, 21, 28, 32].

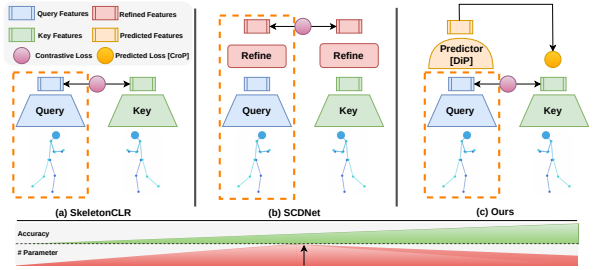

Figure 1. Illustration of the contrastive learning-based skeleton action recognition framework (top) and its accuracy-cost trade-off (bottom) compared to our proposed method **BReSK**.

Although these approaches have achieved promising performance, recent designs often rely on increasingly complex architectural components for feature refinement. For instance, SCDNet [28] introduces a Transformer-based refinement module that enhances feature representations through explicit self-correlation before projecting them into the contrastive space. While such strategies can improve representation quality, they also introduce redundant feature correlations and increase computational cost during both training and inference.

This limitation becomes more critical in real-world activity understanding scenarios. Compared with controlled laboratory datasets, real-world skeleton sequences often contain noisy joints, viewpoint variations, and irregular motion patterns, as illustrated in Fig. 1. Under these conditions, overly complex refinement modules may not necessarily lead to more robust representations, while significantly increasing computational overhead.

To address this limitation, we propose **BReSK**, a simple and efficient self-supervised framework for skeleton representation learning. Instead of explicitly refining features through complex architectural components, BReSK improves representation quality through implicit alignment in the learning objective. The framework combines a bootstrap prediction mechanism with contrastive learning, encouraging stable cross-view representation alignment while maintaining strong instance discrimination. This design strengthens positive pair consistency and reduces feature redundancy in the contrastive space.

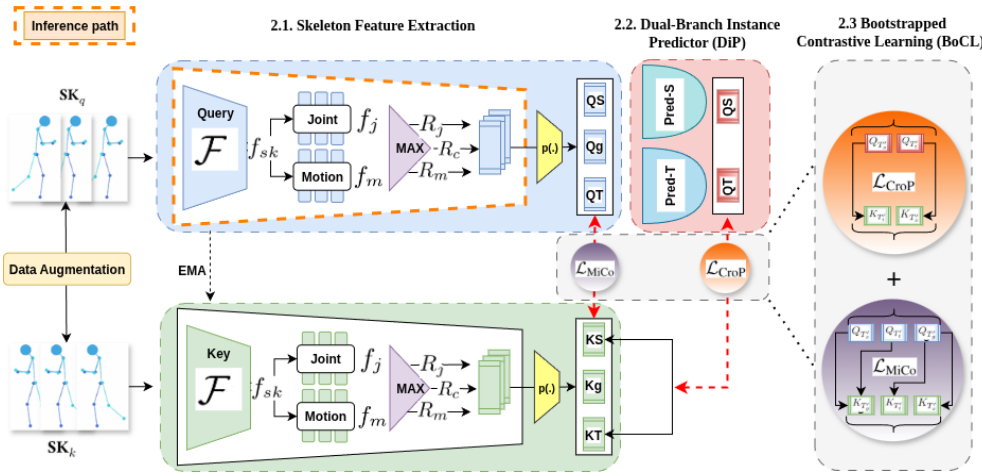

Figure 2. Overview of BReSK framework.

We evaluate BReSK on six benchmark datasets, i.e., Posetics, Toyota SmartHome PennAction, NTU-RGB+D 60/120 and PKU-MMD, covering diverse scenarios including real-world sport and daily living activities. Across these benchmarks, BReSK consistently outperforms recent self-supervised methods. The learned representations remain stable and transferable across datasets, enabling strong performance in downstream tasks such as action recognition and temporal action detection.

More importantly, BReSK achieves these improvements while maintaining significantly lower computational cost. Compared with recent methods relying on complex refinement modules, our framework requires substantially fewer parameters and lower FLOPs while preserving strong representation capability. These results suggest that effective skeleton representation learning can be achieved through carefully designed learning objectives rather than increasingly complex architectures.

In summary, our main contributions are: (i) We propose **BReSK**, a simple and efficient self-supervised framework that combines bootstrap representation alignment with contrastive learning for skeleton-based action understanding. (ii) We introduce an asymmetric instance prediction mechanism that enforces cross-view consistency using spatial and temporal predictors while maintaining stable targets through momentum updates. (iii) We design a hybrid learning objective that jointly optimizes bootstrap prediction alignment and contrastive discrimination, improving representation quality while reducing feature redundancy. (iv) Extensive experiments on six benchmarks demonstrate that BReSK achieves state-of-the-art performance while requiring significantly lower computational cost.

2. Proposed Approach

In this section, we introduce **BreSK**, a self-supervised framework that improves skeleton-based action understanding by combines a bootstrap prediction with contrastive learning.

Overview: As shown in Fig. 2, the proposed framework (BReSK) comprises three core components: (1) **Skeleton Feature Extraction**; (2) **Dual-Branch Instance Predictor**; and (3) **Bootstrapped Contrastive Learning**. Each component is described in detail in the following sections.

2.1. Skeleton Feature Extraction:

The input skeleton sequence $\mathcal{SK} \in \mathbb{R}^{t \times J \times C}$, where t is the sequence length, J is the number of joints, and C denotes joint coordinate dimensions (2D or 3D). Two augmented views—$\mathcal{SK}_q$ (query) and $\mathcal{SK}_k$ (key)—are generated using rotation, flipping, and spatiotemporal masking [28] to improve robustness. Both views are encoded using a CTR-GCN backbone, producing f_{sk_q} and f_{sk_k}, respectively. Inspired by MoCo-v2 [9], the query encoder is updated via backpropagation, while the key encoder is momentum-driven. Both share identical structure but process different augmentations.

Encoder outputs contain tightly coupled spatiotemporal features that are suboptimal for contrastive learning. To disentangle these, we apply MLPs to separate the representations into motion view (f_m) and joints view (f_j) components:

$$f_m, f_j = \text{ReLU}(\text{LayerNorm}(\text{Linear}(f_{sk}))) \quad (1)$$

These intermediate features are projected into the contrastive space via a Max Pooling operation followed by a projection layer in both the branches query and key.

2.2. Dual-Branch Instance Predictor (DiP)

DiP 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

Methods	Linear Evaluation Posetics				Transfer Learning (Pre-trained on Posetics) SmartHome CS		SmartHome CV2		PennAction		Semi-supervised SmartHome CS				PennAction			
	Top-1 (%)		Top-5 (%)		CS (%)		CV2 (%)		Top-1 (%)		5 (%)		10 (%)		5 (%)		10 (%)	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
AimCLR [21] (AAAI 22)	19.2	–	39.3	–	46.6	–	48.3	–	–	–	–	–	–	–	–	–	–	–
CMD [16] (ECCV 22)	20.4	20.4	40.5	40.7	49.0	50.8	52.5	–	89.4	87.4	–	–	–	–	–	–	–	–
HiCLR [4] (AAAI 23)	20.1	–	39.9	–	49.1	–	52.3	–	88.7	–	–	–	–	–	–	–	–	–
HiCo [7] (AAAI 23)	21.3	28.7	42.1	–	54.3	53.3	54.8	56.1	87.6	92.7	34.5	25.5	46.0	34.6	57.2	75.5	74.5	83.3
MAMP [17] (ICCV 23)	–	20.4	–	40.7	–	50.8	–	–	–	87.4	–	–	–	–	–	–	–	–
PCM^3 [33] (ACMMM 23)	20.0	34.5	40.3	58.0	45.3	49.7	46.8	56.6	85.6	94.9	23.1	30.5	30.1	37.1	46.3	78.8	60.9	84.9
UmURL [20] (ACMMM 23)	–	32.7	–	56.5	–	48.1	–	56.9	–	94.8	–	22.7	–	30.6	–	78.1	–	85.9
SCD-Net [28] (AAAI 24)	27.2	–	50.6	–	54.6	–	55.5	–	90.8	–	–	–	–	–	–	–	–	–
ViA [32] (IJCV 24)	20.7	–	40.1	–	49.5	50.5	52.6	–	90.2	–	38.6	–	45.3	–	65.8	–	85.2	–
MacDiff [29] (ECCV 24)	–	23.3	–	42.1	–	51.1	–	56.2	–	91.9	–	–	–	–	–	–	–	–
USDRL [27] (AAAI 25)	25.9	34.6	48.6	58.5	55.0	49.6	53.9	56.7	89.6	95.2	37.1	34.3	43.6	40.3	58.1	73.9	70.3	86.4
BReSK (Our)	**29.8**	**36.9**	**52.7**	**61.0**	**56.5**	**53.9**	**56.9**	**57.8**	**91.9**	**95.8**	**39.7**	**37.4**	**48.3**	**46.5**	**69.5**	**80.4**	**86.1**	**89.7**

Table 1. **2D vs 3D Comparison** across Linear, Transfer, and Semi-supervised settings on Posetics, SmartHome, and PennAction.

Method	NTU-RGB+D 60		NTU-RGB+D 120	
	CS(%)	CV(%)	CS(%)	CSet(%)
CrosSCLR [13](CVPR21)	72.9	79.9	-	-
HiCLR [4](AAAI23)	80.4	85.5	70.0	70.4
UmURL [20](ACMMM 23)	82.3	89.8	73.5	74.3
HiCo [7](AAAI 23)	81.1	88.6	72.8	74.1
ActCLR [14](CVPR 23)	80.9	86.7	69.0	70.5
ViA [32](IJCV 24)	78.1	85.8	69.2	66.9
SCD-Net [28] (AAAI 24)	86.6	91.7	76.9	80.1
ACA2Net [2](TCSVT 25)	86.0	89.6	-	-
ActCLR+ [15](TPAMI 25)	82.3	88.2	70.9	73.2
Heterogeneous [23] (CVPR 25)	80.2	88.0	70.7	73.5
USDRL [27] (AAAI 25)	84.2	90.8	76.0	76.9
USDRL+ [24] (TPAMI 25)	85.8	91.8	77.5	78.8
PCM^3++ [34] (IJCV 26)	84.8	91.0	76.7	-
BReSK (Our)	**87.3**	**92.0**	**79.1**	**80.1**

Table 2. **Action Recognition**- *Linear Evaluation results*- for Lab-setting datasets (NTU-RGB+D 60 and NTU-RGB+D 120)

Method	mAP@tIoU (%)		
	0.1	0.3	0.5
AimCLR [21]	43.9	-	35.1
HiCo-Transformer [7]	32.5	31.8	28.6
HiCo-GRU [7]	50.1	48.6	44.3
SkeAttnCLR [10]	48.5	-	41.7
LAC [31]	55.2	-	-
SCDNet [28]	65.6	64.7	58.5
BReSK (Our)	**69.5**	**68.5**	**64.0**

Table 3. **Action Detection**- *Linear Evaluation results*- for Lab-setting datasets (PKU-MMD I).

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:

$$f_{i\to j}(q_i) = W_2^{(i\to j)}\,\sigma\Big(\mathrm{LN}\Big(W_1^{(i\to j)} q_i\Big)\Big), \qquad (2)$$

where $W_1 \in \mathbb{R}^{d_h\times d}$ and $W_2 \in \mathbb{R}^{d\times d_h}$ are learnable parameters, $\mathrm{LN}(\cdot)$ denotes Layer Normalization, and $\sigma(\cdot)$ is the ReLU activation function. In practice, we employ two cross-view predictors:

$$\hat{q}_s = f_{t\to s}(q_t), \quad \hat{q}_t = f_{s\to t}(q_s), \qquad (3)$$

where $\hat{q}_s$ and $\hat{q}_t$ denote the predicted representations in the spatial and temporal spaces, respectively. By explicitly learning mappings across views, the proposed predictors encourage the model to capture complementary spatial-temporal semantics and improve representation alignment.

2.3. Bootstrapped Contrastive Learning (BoCL)

BoCL jointly optimizes 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.

Objective 1: Cross-Branch Prediction Alignment. Our Dual-Branch Instance Predictor (DiP) learns to match the output of a momentum-updated teacher (key) network using stop-gradient targets and a predictor-based alignment. To enforce cross-view consistency, we apply a cosine similarity loss between predicted features and their corresponding target representations.

BReSK consists of two branches: (i) a *query branch*, optimized via back-propagation and equipped with cross-view predictors, and (ii) a *key branch*, implemented as a momentum encoder updated using an exponential moving average (EMA) of the query parameters. The key branch provides stable targets and is detached from gradient computation.

Let k_j denote the target representation from view j, produced by the EMA encoder, and let $\hat{q}_j = f_{i\to j}(q_i)$ be the predicted representation from the query branch. We normalize both vectors as:

$$\tilde{q}_j = \frac{\hat{q}_j}{\|\hat{q}_j\|_2}, \quad \tilde{k}_j = \frac{\mathrm{sg}(k_j)}{\|k_j\|_2}, \qquad (4)$$

where $\mathrm{sg}(\cdot)$ is the stop-gradient operator, ensuring gradients flow only through the query branch. The cross-view alignment loss is defined as:

$$\mathcal{L}_{i\to j} = \frac{1}{N}\sum_{n=1}^{N}\Big(1 - \tilde{q}_j^{(n)}\cdot\tilde{k}_j^{(n)}\Big), \qquad (5)$$

which minimizes the negative cosine similarity between predictions and targets. This encourages the query branch to align its representations with the EMA teacher across different views. Finally, the objective is symmetrized across both directions:

$$\mathcal{L}_{\text{CroP}} = \mathcal{L}_{t\to s} + \mathcal{L}_{s\to t}. \qquad (6)$$

Objective 2: Contrastive Learning. This objective enhances discrimination in the embedding space by mixing spatial, temporal, and global views from different branches. Following [28], the MiCo loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{MiCo}} = {} & \mathcal{L}_{\text{info}}(Q_j, K_c) + \mathcal{L}_{\text{info}}(Q_m, K_c) \\ & + \mathcal{L}_{\text{info}}(Q_c, K_j) + \mathcal{L}_{\text{info}}(Q_c, K_m), \end{aligned} \qquad (7)$$

CroP loss form	loss terms $(Query \rightarrow Key)$	acc
Instance	$(S \rightarrow S) + (T \rightarrow T) + (G \rightarrow G)$	86.6
Sub instance	$(S \rightarrow S) + (T \rightarrow T)$	86.4
Mixed	$(S \rightarrow G) + (T \rightarrow G)$	86.3
Cross	$(S \rightarrow T) + (T \rightarrow S)$	**87.3**

Table 4. Prediction loss forms

where Q and K represent query and key features, and $\mathcal{L}_{\text{info}}$ is the InfoNCE loss [22].

Final Loss. The total self-supervised loss combines both objectives, $\mathcal{L} = \mathcal{L}_{\text{CroP}} + \mathcal{L}_{\text{MiCo}}$. This complementary strategy ensures that representations are internally consistent and externally discriminative, a key for robust generalization in downstream tasks.

3. Experiments and Analysis

We conduct extensive experiments comparing with state-of-the-art (SoTA) SSL methods to validate the effectiveness and generalization of BReSK across diverse settings. Evaluation protocols, additional implementation details, dataset descriptions, experimental details, and results are provided in the **Appendix**.

3.1. Comparison with State-of-the-Art

Action Recognition - Linear Evaluation. As reported in Table 1 (**blue**), BReSK achieves SoTA performance across all real-world benchmarks, outperforming ViA [32] by **+9.1%** (Top-1) and **+12.6%** (Top-5) on **Posetics**. On controlled lab benchmarks (Table 2), BReSK consistently surpasses strong prior methods such as SCD-Net [28] on **NTU-RGB+D 60/120**, achieving a **+2.2%** improvement on NTU-120 under the Cross-Subject protocol using the joint modality.

Action Recognition - Transfer Learning Evaluation. We further evaluate representation transferability across datasets. In the real-world setting, we pre-train on **Posetics** and transfer to unseen datasets using linear evaluation (Table 1, **green**). BReSK outperforms prior state-of-the-art methods by **+1.5%** under the Cross-Subject protocol and **+1.4%** under Cross-View2 on **SmartHome**, demonstrating strong robustness to domain shifts and camera variations.

Action Recognition - Semi-supervised Evaluation. We evaluate BReSK under limited labeled data by fine-tuning with 5% and 10% of annotations on **SmartHome** and **Penn-Action** (Table 1, **yellow**). BReSK maintains strong performance under these low-label settings, demonstrating that slot-structured learning effectively captures transferable motion priors. Extending this evaluation to 3D skeletons (BReSK-3D), we observe consistent improvements across datasets, particularly for general actions in Posetics and sports actions in PennAction.

Action Detection – Linear Evaluation. We further evaluate BReSK on temporal localization using the untrimmed **PKU-MMD I** dataset. Following [28], the encoder is pre-trained on NTU-60 (Cross-Subject) and a linear classifier is trained to predict frame-level actions. As shown in Table 3, BReSK achieves significant mAP improvements over recent self-supervised baselines [7, 31], demonstrating that its slot-structured motion representation and unified discriminative reconstruction effectively capture temporal dynamics and enhance precise action localization in untrimmed sequences.

Framework	Contrastive	DiP	CroP	NTU60-CS (%)	NTU120-CS (%)
Baseline	✓(NCE)	✗	✗	83.5	75.0
BReSK (Ours)	✓(NCE)	✓	✓	**85.8**	**76.7**
Baseline	✓(MiCo)	✗	✗	84.0	75.8
-	✓	✗	✓	86.7	76.9
-	✓	✓	✗	87.0	77.5
BReSK (Ours)	✓(MiCo)	✓	✓	**87.3**	**79.1**

Table 5. Ablation study of different components.

Model	Inference		Training			NTU120-CS (%)
	Params	GFlops	Params	Time (h)	Memory	
H-Transformer [5]	>100 M	118.6	—	—	—	—
GL-Transformer [11]	214 M	—	214 M	18	—	66.0
PCM3 [33]	—	—	103 M	15	394.66 MB	76.5
SCDNet [28]	83.26 M	7.14	184 M	17	702.04 MB	76.9
USDRL [27]	94.86 M	—	95 M	14	362.13 MB	76.0
BReSK (Ours)	**22.03 M**	**3.75**	**83 M**	**11**	**318.74 MB**	**79.1**

Table 6. Computation cost on NTU-RGB+D 120 CS during inference(blue) and training process (green).

3.2. Ablation Study

Component-wise Analysis of BReSK: Table 5 shows that both DiP and CroP consistently improve over the baseline across different contrastive objectives. While each component individually provides noticeable gains, their combination leads to the best performance, indicating a clear complementary effect. This is further validated under the stronger MiCo setting, where the full BReSK model achieves the highest accuracy.

Effect of Cross-Space Feature Alignment: Table 4 shows the importance of cross-space feature alignment in improving representation quality. While instance-level and sub-instance losses based on intra-space mappings (e.g., $S \rightarrow S$, $T \rightarrow T$) yield comparable performance, they are consistently outperformed by formulations that encourage inter-space interactions. Notably, the cross loss ($S \rightarrow T, T \rightarrow S$) achieves the highest accuracy (87.3%), demonstrating that explicitly modeling complementary relationships between spatial and temporal features leads to more discriminative representations.

Computation Cost: Table 6 shows that BReSK achieves SOTA performance across all evaluation protocols, while simultaneously providing a significant reduction in computational cost, with up to $\sim$2× fewer inference parameters and over 3× lower GFLOPs compared to existing methods.

4. Conclusion

We presented BReSK, a self-supervised framework that leverages bootstrap prediction with contrastive learning for skeleton action recognition. Our design improves cross-view representation consistency while maintaining strong instance discrimination. Experiments on multiple benchmarks show that BReSK consistently outperforms existing methods across diverse settings with lower computational cost.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *ICCV*, 2021. 1

[2] 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. 3

[3] 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 *ICCV*, 2021. 1

[4] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N. Metaxas. Hierarchically self-supervised transformer for human skeleton representation learning. In *ECCV*, 2022. 3

[5] Yi-Bin Cheng, Xipeng Chen, Junhong Chen, Pengxu Wei, Dongyu Zhang, and Liang Lin. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 4

[6] Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, 2024. 1

[7] 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. 3, 4

[8] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 2021. 1

[9] 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

[10] 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. 3

[11] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 2022. 4

[12] 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. 1

[13] 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

[14] 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

[15] 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. 3

[16] 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. 1, 3

[17] 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. 3

[18] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 1

[19] Chen Sun, Arsha Nagrani, Yonglong Tian, and Cordelia Schmid. Composable augmentation encoding for video representation learning. In *ICCV*, 2021. 1

[20] 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

[21] 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. 1, 3

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 4

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

[24] 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

[25] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 1

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. 1

[27] 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. 3, 4

[28] 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, 4

[29] Lehong Wu, Lilang Lin, Jiahang Zhang, Yiyang Ma, and Jiaying Liu. Macdiff: Unified skeleton modeling with masked conditional diffusion. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[30] 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

[31] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac - latent action composition for skeleton-based action segmentation. In *ICCV*, 2023. 3, 4

[32] 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

[33] 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, 4

[34] Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. Self-supervised skeleton-based action representation learning: A benchmark and beyond. *IJCV*, 2026. 3

[35] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023. 1