Bayesian Contrastive Learning with Manifold Regularization for Self-Supervised Skeleton Based Action Recognition

Lilang Lin, Jiahang Zhang, Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

2023 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-5109-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCAS46773.2023.10181797

***Abstract*—In this paper, we address skeleton-based action recognition under the self-supervised setting. We propose a novel framework *Bayesian Contrastive Learning with Manifold Regularization* (*BCLR*). In Bayesian contrastive learning, we employ Monte Carlo Dropout sampling on the adjacency matrix of the skeleton data to obtain positive/negative samples for model robustness. A novel entropy-based memory bank updating strategy is further proposed to take full advantage of hard negative samples for better separability. The feature manifold regularization, including projection-based data reconstruction and similarity-based feature decoupling, on the other hand, is designed to extract comprehensive information to avoid overfitting and increase feature diversity to prevent a collapse of the model. With Bayesian contrastive learning and feature manifold regularization, our model learns stronger and more discriminative features. Extensive experiments on NTU RGB+D and PKUMMD show that the proposed method achieves remarkable action recognition performance.**

***Index Terms*—skeleton based action recognition, contrastive learning, bayesian neural network, self-supervised learning**

I. Introduction

Action recognition is a critical and challenging field in computer vision and has wide application in human-computer interaction, video surveillance, virtual reality, *etc*. Human skeletons describe human behaviors with skeleton joints using the 3D coordinate locations, and have attracted increasing attention in recent years [1–6]. It can be easily captured by depth sensors or extracted from other modalities with mature algorithms [7]. Compared with RGB videos or depth data, skeletons are lightweight, privacy-preserving, and robust to views, appearances, and backgrounds. Moreover, skeletons are higher-level feature representations of human motion, which are easier for analysis and more discriminative for action recognition.

Conventional skeleton-based action recognition methods [8–10] require numerous labeled training data and are of limited flexibility for practical application. To get rid of the reliance on full supervision, Zheng *et al*. [11] first proposed to effectively learn skeleton representations from unlabeled data. Since then, researchers have been made great efforts for self-supervised skeleton-based action recognition [6, 12, 13]. Current methods can be classified into two categories: reconstruction-based and contrastive learning. *Reconstruction-based* methods [6] apply encoder-decoder structures, where the encoder extracts features from the original or part of the skeleton data, and the decoder reconstructs the skeleton based on the extracted features. *Contrastive learning* [12–14] employs data transformation to generate positive/negative samples and narrows the distance between positive samples while increasing the distance between negative samples for intra-class tightness and inter-class separability.

However, for reconstruction-based methods, there is a severe task gap between generation and recognition, which can degrade the performance of downstream tasks. And for contrastive learning, first, there is a lack of transformation designed specifically for skeletons, while the performance of contrastive learning is greatly affected by data transformation. Second, skeleton data, as a high-level representation, is very sensitive to data transformation. Strong data transformation tends to lose important information, while weak data transformation is not effective. This property adds to the difficulty of method design. Moreover, contrastive learning often has shortcuts, which can easily lead to feature overfitting and model collapse [15]. More constraints need to be provided in the feature space to facilitate contrastive learning.

In order to address the aforementioned issues, we introduce *Bayesian Contrastive Learning with manifold Regularization*, which exploits model transformation to generate positive/negative samples, and extracts more separable representations through the constraints of the feature space. We propose two modules, Bayesian contrastive learning and feature manifold regularization. Bayesian contrastive learning transforms the model based on Monte Carlo dropout, and employs the transformed model to extract different features as positive pairs. By training with dropout, the model improves the robustness to disturbance. Besides, to make full use of negative samples, an entropy-based memory bank updating strategy is proposed, which preferentially selects negative samples with small entropy to be replaced. Thus, the negative samples with large entropy, which are harder negative samples, are retained and make the model learn a more separable feature space. Feature manifold regularization applies projection-based data reconstruction to extract comprehensive information to avoid overfitting for contrastive learning and similarity-based feature

*Corresponding author. This work was supported by the National Natural Science Foundation of China under Contract No.62172020, and a research achievement of Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

decoupling for increasing the diversity of features to prevent degenerate solutions.

In summary, our contributions include the following: 1) We construct a self-supervised learning mechanism *Bayesian Contrastive Learning with manifold Regularization*, which employs contrastive learning to extract robust and separable feature representations and applies auxiliary tasks to constrain the features for better generalization ability. 2) We propose a Bayesian contrastive learning paradigm to obtain stronger feature representation capacity. Based on model transformation, we reduce epistemic uncertainty for better model robustness. And we apply entropy to improve the updating strategy of memory bank. This strategy makes full use of negative samples to obtain the separable feature space. 3) We design feature manifold regularization. Through projection-based data reconstruction and similarity-based feature decoupling, The model avoids overfitting to shortcuts and collapse solutions.

The rest of the paper is organized as follows. Sec. II introduces the proposed bayesian contrastive learning method. Experimental results are shown in Sec. III and concluding remarks are provided in Sec. IV.

II. Bayesian Contrastive Learning Method

The proposed self-supervised learning method is interpreted in this section. In summary, we deploy Bayesian contrastive learning to extract feature representations and manifold regularization to boost the feature space.

A. Bayesian Deep Learning for Skeleton Based Action Recognition

Previous work pointed out that there is a strong correlation between the performance of contrastive learning and data transformation. Therefore, the research on data transformation has been quite rich. However, most of these previous methods are designed for RGB image data, lack of generalizibility for skeleton data. And because high-level representation of information like skeleton is sensitive to data transformation, a too strong strategy of transformation may lose too much information, and a too weak strategy can not be efficient enough for feature learning. So it is non-trivial to find the optimal data transformation.

Therefore, our attention move from transformation of the data to that of the model, *i.e.* generating positive and negative samples by transforming the model weights. To achieve model transformation in a neural network, we replace deterministic weight parameters of the network with distributions over these parameters, which turns the neural network into a Bayesian neural network (BNN). We employ dropout variational inference to sample the parameters from the distribution to obtain the transformed model. This inference is done by training a model with dropout before every layer to sample from the approximate posterior (stochastic forward passes, referred to Monte Carlo dropout [16]). And based on the BNN, we estimate the aleatoric uncertainty and epistemic uncertainty [16] to assist training.

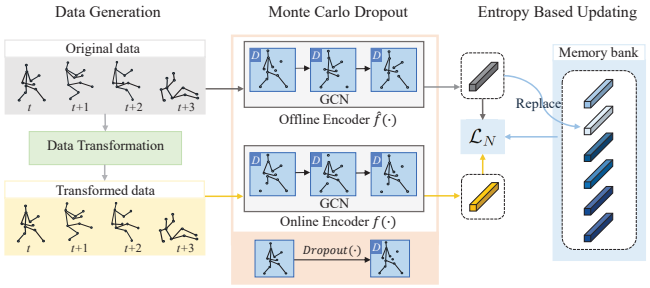

Fig. 1. The composition of Bayesian contrastive learning. Yellow and gray arrows are applied to indicate the pipeline for the online encoder $f(\cdot)$ branch and the offline encoder $\hat{f}(\cdot)$ branch, respectively. On each forward pass, we employ Monte Carlo Dropout to sample the graph from the Bernoulli distribution. The blue arrows show the pipeline of the entropy based memory bank updating strategy. The features with darker colors in the memory bank represent larger entropy. The feature embedding with the smallest entropy is replaced when updating.

As shown in Fig. 1, our pipeline utilizes the encoder $f(\cdot)$ and adopts a projection head for contrastive learning. More details about the Bayesian contrastive learning are provided as follows.

Bayesian Contrastive Learning. Based on BNN, we propose *Monte Carlo dropout* as a model transformation and employ uncertainty to update the memory bank adaptively.

• **Monte Carlo Dropout**: BNN assumes that the parameters of the model come from a distribution, and each forward pass samples the model from the distribution to obtain the output $\mathbf{h}^i = f^{\widehat{\mathbf{W}}}(\mathbf{x}^i)$ with $\widehat{\mathbf{W}} \sim q(\mathbf{W})$, where $\mathbf{W}$ is the parameters to be optimized and $q(\cdot)$ is the distribution function. For Monte Carlo dropout, we adopt Bernoulli distribution for sampling. Specifically, we only transform the adjacency matrix in the graph convolutional network (GCN), which makes the network more robust to modeling the relationship between nodes. The transformed model can be expressed as:

$$\mathbf{h}_{l+1}^i = \sigma(\mathbf{D}^{-\frac{1}{2}}(\mathbf{B} \odot \mathbf{A})\mathbf{D}^{-\frac{1}{2}}\mathbf{h}_l^i\mathbf{W}_l), \tag{1}$$

where $\odot$ indicates element-wise multiplication and $\mathbf{B} \sim \frac{1}{p}\mathbf{Bernoulli}(p)$. The matrix $\mathbf{A}$ is the adjacency matrix.

Sampling from a model distribution is equivalent to an ensemble of multiple models. Using features extracted from different transformed models for contrastive learning can reduce epistemic uncertainty, which captures uncertainty in the model parameters. Thus, the model can learn to extract more consistent feature representations and becomes more robust to noisy data.

• **Entropy Based Memory Bank Updating Strategy**: Previous works treat each negative sample equally and perform the first-in, first-out updating strategy, ignoring the differences of the negative samples. To make more use of harder samples for representation learning, we apply aleatoric uncertainty to quantitatively measure the difficulty of negative samples, and employ it as an indicator to preferentially replace samples with low uncertainty and retain samples with high uncertainty. Heteroscedastic aleatoric uncertainty captures noise inherent in the observations, with some samples potentially having

more noisy outputs than others. These samples are regarded as harder negative samples.

To capture aleatoric uncertainty in contrastive learning, we compute the entropy in contrastive learning for each sample. It can be measured as follows:

$$H(\hat{\mathbf{s}}^i) = -\sum_{j=1}^{M} \mathbf{p}^j \log \mathbf{p}^j, \mathbf{p}^j = \frac{\exp(K(\hat{\mathbf{s}}^i, \mathbf{m}^j))}{\sum_{k=1}^{M} \exp(K(\hat{\mathbf{s}}^i, \mathbf{m}^k))}, \quad (2)$$

where $\mathbf{m}^j \in \mathbf{M}$. The uncertainty of the sample can be measured by entropy *i.e.* samples with higher entropy are more easily confused and therefore more uncertain. These samples are treated as harder negatives and kept in memory bank updating. And the samples with the smallest entropy are replaced by new negative samples.

However, considering this entropy is an inaccurate estimate of aleatoric uncertainty due to the large model uncertainty in the early training stage, we perform a two-stage training strategy: 1) first-in, first-out strategy for memory bank updating. 2) then the model can supply high-confidence aleatoric uncertainty estimation, the strategy is replaced with *Entropy Based Memory Bank Updating Strategy*.

B. Feature Manifold Regularization

The fore-mentioned design learns a compact and separable feature space by contrastive learning. However, contrastive learning faces problems such as overfitting and pseudo-negative samples, which make the learned features not well adaptive to downstream tasks. In this section, we introduce two agent tasks, projection-based data reconstruction and similarity-based feature decoupling, to boost the representation learning.

Projection-Based Data Reconstruction. Previous works often employ original data for reconstruction. Instead, we apply random projection for manifold projection, which is simple and computationally efficient. Besides, it can approximately preserve the paired distance between any two samples in the data set according to Johnson–Lindenstrauss lemma [17].

After sampling the random projection $\mathcal{A}$, we project the data $\mathbf{x}^i$ into the feature space $\tilde{\mathbf{h}}^i$. Finally, a regression head $g_r(\cdot)$ takes in the features $\mathbf{h}^i$ extracted by the encoder $f(\cdot)$ to reconstruct the projection $\tilde{\mathbf{h}}^i$. Data reconstruction enables the model to extract more comprehensive information without losing important information due to overfitting in contrastive learning.

Similarity-Based Feature Decoupling. In order to satisfy the diversity assumption, we achieve feature decoupling by reducing the similarity between each other. Specifically, we apply both online similarity reduction and offline similarity reduction to exploit both the data within a batch and the features in the memory bank.

• **Online Similarity Reduction**: To reduce the similarity of features in a batch for each sample, we consider reducing the concentration on the spherical manifold by employing a projection head $g_{on}(\cdot)$ to project and normalize the output features $\mathbf{h}^i$ to unit vectors $\mathbf{c}^i$ on the hypersphere. After sampling a batch $\mathbf{B}$, we compute the center of mass vector $\bar{\mathbf{c}} = \frac{1}{|\mathbf{B}|} \sum_{\mathbf{c}^i \in \mathbf{B}} \mathbf{c}^i$. The concentration is the length of the center of mass vector, given by $\bar{\mathbf{R}} = \sqrt{\bar{\mathbf{c}}^T \bar{\mathbf{c}}}$.

If the features are tightly clustered, $\bar{\mathbf{R}}$ will be almost 1. If features are widely dispersed, $\bar{\mathbf{R}}$ will be almost 0. We make the feature distribution more uniform by reducing $\bar{\mathbf{R}}$.

• **Offline Similarity Reduction**: To exploit the features in the memory bank $\mathbf{M}$, we randomly sample in the bank $\mathbf{M}$ to construct a decodable information bottleneck (DIB) [18]. In detail, we first randomly sample a feature $\mathbf{m}^r$ from the memory bank $\mathbf{M}$. Then we train a projection network $g_{off}(\cdot)$ to employ the input feature $\mathbf{h}^i$ to increase the cosine similarity with $\mathbf{m}^r$. In training, we reverse the gradients of encoder so that the features extracted by the encoder $f(\cdot)$ are not easy to be used to predict $\mathbf{m}^r$.

Resort to this module, the similarity between different features is reduced with the diversity increased. The task reduces the predictability between them and the redundancy of features, which makes them more independent.

TABLE I
USUPERVISED LEARNING RESULTS ON PKUMMD DATASET.

Models	PKU I (%)	PKU II (%)
MS^{2}L [12]	64.8	27.6
P&C [6]	59.9	25.5
3s-AimCLR [24]	-	38.5
ISC [25]	80.9	36.0
3s-CrosSCLR [14]	84.9	32.9
BCLR	**85.6**	**44.5**

III. EXPERIMENT RESULTS

To train the network, all skeleton sequences are temporally down-sampled to 50 frames. The encoder $f(\cdot)$ is based on *ST-GCN* [19] with hidden units of size 256. Adam optimizer [20] is applied for training of 300 epochs. We conduct the experiments on following datasets:

• **NTU RGB+D Dataset 60 (NTU 60) [21]** This is a large-scale dataset which contains 56,578 videos with 60 action labels and 25 joints for each body, including interactions with pairs and individual activities.

• **NTU RGB+D Dataset 120 (NTU 120) [22]** This is an extension to NTU 60 and the largest dataset for action recognition, which contains 114,480 videos with 120 action labels. Actions are captured with 106 subjects with multiple settings using 32 different setups.

• **PKU Multi-Modality Dataset (PKUMMD) [23]** The actions are organized into 52 categories and include almost 20,000 instances in PKUMMD. The PKUMMD is divided into part I and part II. Part II provides more challenging data with large view variation. We evaluate the model on the cross-subject (xsub) protocol.

A. Evaluation and Comparison

In this part, we compare our method with other methods under unsupervised, semi-supervised and supervised settings.

TABLE II
UNSUPERVISED LEARNING RESULTS ON NTU DATASET.

Model	NTU 60 (%)		NTU 120 (%)	
	xsub	xview	xsub	xset
MS^{2}L [12]	-	52.5	-	-
AS-CAL [13]	64.8	58.5	49.2	48.6
P&C [6]	59.3	56.1	44.1	41.4
SeBiReNet [26]	79.7	-	-	-
AimCLR [24]	79.7	74.3	-	-
3s-Colorization [27]	83.1	75.2	-	-
ISC [25]	78.6	76.3	**67.1**	**67.9**
3s-CrosSCLR [14]	83.4	**77.8**	66.7	**67.9**
BCLR	**83.9**	**77.8**	66.8	67.5

TABLE III
SEMI-SUPERVISED LEARNING RESULTS ON NTU 60 DATASET.

Models	1%		10%	
	xview	xsub	xview	xsub
ASSL [28]	-	-	69.8	64.3
MCC [29]	-	-	59.9	55.6
Colorization [27]	-	-	73.3	66.1
ISC [25]	38.1	35.7	72.5	65.9
BCLR	**42.2**	**37.6**	**73.4**	**68.8**

Unsupervised Approaches. In unsupervised learning, a linear classifier $\phi(\cdot)$ is applied to the pre-trained fixed encoder $f(\cdot)$ to classify the extracted features. Compared with other methods in Table I and II, our model shows superiority on these datasets. We argue that the transformation that 3s-CrosSCLR [14] and ISC [25] design in contrastive learning task is handcrafted, which makes it easier for the model to overfit the handcrafted transformation. On the contrary, our method adopts Bayesian contrastive learning and the negative samples are adaptively updated due to our specific design. Moreover, through the feature space regularization, the features extracted by our method are more suitable for downstream tasks.

Semi-Supervised Approaches. In semi-supervised learning, the encoder $f(\cdot)$ is pretrained with full data, and then fine-tuned with the classifier $\phi(\cdot)$ with with randomly sampled 1%, 10% of the training data. In Table III, our method improves the accuracy considerably and performs better than the state-of-the-art methods, especially with smaller training data.

Supervised Approaches. In the supervised learning setting, after pretraining on the encoder $f(\cdot)$, we fine-tune the encoder $f(\cdot)$ and classifier $\phi(\cdot)$ on the downstream task. The results in Table IV confirm that our method extracts the information demanded by downstream tasks and can better benefit action recognition. In comparison with state-of-the-art supervised learning methods, our model achieves better performance.

TABLE IV
SUPERVISED LEARNING RESULTS ON PKUMMD DATASET.

Models	PKU I (%)	PKU II (%)
MS^{2}L [12]	83.4	42.4
ISC [25]	84.0	40.2
3s-CrosSCLR [14]	91.1	50.8
BCLR	**91.4**	**55.0**

TABLE V
ANALYSIS OF BAYESIAN CONTRASTIVE LEARNING ON PKUMMD I DATASET WITH UNSUPERVISED LEARNING APPROACHES.

Ablation	Configuration	PKUMMD I (%)
Probability of Dropout	0	84.3
	0.125	85.6
	0.25	51.5
	0.5	50.2
Model Transformation	$\mathbf{B} \odot \mathbf{A}$	85.6
	$\mathbf{B} \odot \mathbf{W}_l$	84.7
	$\mathbf{B} \odot \mathbf{h}_l^i$	84.5
Updating Strategy	Queue-Based	84.7
	Entropy-Based	85.6

TABLE VI
ANALYSIS OF MODULE COMBINATION ON PKUMMD II DATASET WITH UNSUPERVISED LEARNING APPROACHES.

Module			PKUMMD II (%)
BCL	PDR	SFD	
			37.7
✓			38.7
	✓		41.1
		✓	39.5
✓	✓		42.5
✓	✓	✓	44.5

B. Ablation Study

Next, we conduct ablation experiments to give a more detailed analysis of our proposed approach.

In Table V, we first explore the effect of different model transformation settings on the downstream task. As the probability of dropout increases, the accuracy first increases and then decreases. In the absence of model augmentation, contrastive learning is too easy and thus difficult to learn meaningful feature representations. However, too strong a model transformation will result in the loss of too much information. Different transformation positions also have an impact on contrastive learning. Transforming the graph is better than transforming the weights and the features. Finally, experiments on the memory bank updating strategy also demonstrate that our entropy-based strategy can provide harder negative examples to learn a more separable feature space.

The ablation studies on different modules are displayed in Table VI. Contrastive learning provides the most improvement. Feature regularization constraints further optimize the feature space, making features more suitable for generalization to downstream tasks. The combination of multiple tasks achieves the best performance.

IV. CONCLUSIONS

In this work, we propose a self-supervised learning method *Bayesian Contrastive Learning with manifold Regularization* for skeleton-based action recognition. Based on Bayesian deep learning, we propose Bayesian contrastive learning and feature manifold regularization. Specifically, we exploit Monte Carlo Dropout sampling to obtain positive/negative samples. Meanwhile, multiple agent tasks including projection-based data reconstruction and similarity-based feature decoupling to learn a more robust and generalized feature space.

REFERENCES

[1] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[2] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[3] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[4] Y. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. ACM Int'l Conference on Multimedia*, 2020.

[5] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching." in *Proc. AAAI Conference on Artificial Intelligence*, 2020.

[6] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[7] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[8] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[10] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. Int'l Conference on Computer Vision*, 2021.

[11] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[12] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *Proc. ACM Int'l Conference on Multimedia*, 2020.

[13] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Information Sciences*, vol. 569, Aug 2021.

[14] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

[15] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, "Can contrastive learning avoid shortcut solutions?" *Proc. Advances in Neural Information Processing Systems*, vol. 34, 2021.

[16] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Proc. Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] P. Frankl and H. Maehara, "The johnson-lindenstrauss lemma and the sphericity of some graphs," *J. Comb. Theory, Ser. B*, vol. 44, no. 3, pp. 355–362, 1988.

[18] Y. Dubois, D. Kiela, D. J. Schwab, and R. Vedantam, "Learning optimal representations with the decodable information bottleneck," in *Proc. Advances in Neural Information Processing Systems*, 2020.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[20] W. K. Newey, "Adaptive estimation of regression models via moment restrictions," *Journal of Econometrics*, vol. 38, no. 3, pp. 301–339, 1988.

[21] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[23] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 16, no. 2, pp. 41:1–41:24, 2020.

[24] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[25] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3D action representation learning," in *Proc. ACM Int'l Conference on Multimedia*, 2021.

[26] Q. Nie, Z. Liu, and Y. Liu, "Unsupervised 3D human pose representation with viewpoint and pose disentanglement," in *Proc. European Conference on Computer Vision*. Springer, 2020.

[27] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3D action representation learning," in *Proc. Int'l Conference on Computer Vision*, 2021.

[28] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3D action recognition," *Proc. European Conference on Computer Vision*, 2020.

[29] Y. Su, G. Lin, and Q. Wu, "Self-supervised 3D skeleton action representation learning with motion consistency and continuity," in *Proc. Int'l Conference on Computer Vision*, 2021.