Multi-Scale Skeleton Action Recognition Based on Self-Supervision

Zesen Cai, Lisi Mo, Chenxi Li, Ruiting Dai*
University of Electronic Science and Technology of China
Chengdu, China
{zesen.cai, chenxi.li}@std.uestc.edu.cn, {morris, rtdai}@uestc.edu.cn

***Abstract*—Self-supervised learning (SSL) provides a scalable and annotation-efficient paradigm for action recognition. However, existing approaches often struggle to bridge the semantic gap between fragmented local motions and global action contexts, as they predominantly rely on instance-level discrimination while overlooking the multi-level hierarchical nature and long-range spatiotemporal dependencies inherent in human dynamics. To address this, we propose Multi-Scale Self-Distilled Learning (MS-Distill), a self-supervised framework that employs student-teacher networks with a multi-view consistency objective. Unlike traditional approaches, MS-Distill eliminates the reliance on negative samples by aligning representations across various temporal resolutions and spatial granularities. Furthermore, we introduce a Global-Local Spatiotemporal Encoder that models interactions among arbitrary spatiotemporal points, yielding highly discriminative unified embeddings that capture both fine-grained motion dynamics and holistic, long-range action patterns. Extensive experiments on the NTU-RGB+D 60 and 120 datasets demonstrate the superiority of MS-Distill. Notably, it achieves a significant 3.8% gain in KNN accuracy on the NTU-120 xset protocol with minimal pre-training overhead, while exhibiting strong transferability and scalability to downstream tasks.**

***Index Terms*—Self-supervised learning, Skeleton action recognition, Multi-scale representation learning, Self-distillation, Spatiotemporal Transformer**

I. Introduction

Recognizing human actions is a fundamental challenge in machine perception with broad impact on applications such as human–computer interaction [1], augmented reality [2], and intelligent surveillance [3]. While appearance-based approaches have achieved significant progress, their robustness is often undermined by background clutter, illumination variations, and occlusions. This has motivated increasing interest in structured temporal data—particularly human skeletons—which provides a compact and noise-resistant representation of motion. Advances in depth sensing [4], [5] and pose estimation [6], [7] further make high-quality skeleton sequences readily available, enabling scalable research on human actions.

Current approaches in action recognition remain dominated by fully supervised paradigms [8]–[10], whose dependence on large-scale annotated datasets limits their applicability in low-resource or cross-domain settings. Self-supervised learning (SSL) offers a scalable alternative by mining supervisory signals from unlabeled sequences. Early SSL methods [11], [12] relied on proxy tasks such as reconstruction or future-frame prediction, which capture only limited discriminative structure. Contrastive learning [13], [14] has emerged as the leading paradigm, exploiting positive–negative pair optimization to learn structured, generalizable representations.

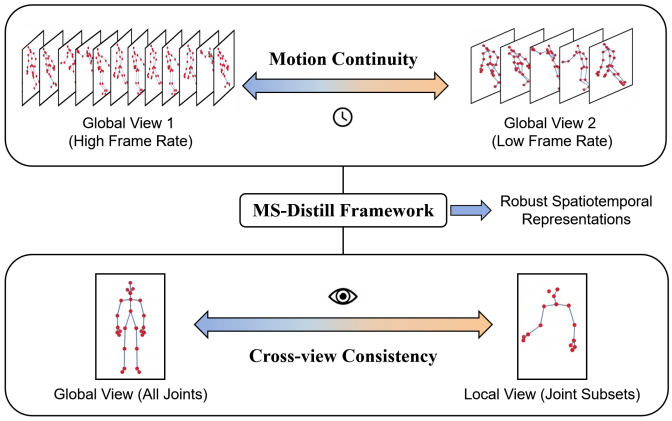

Fig. 1. Conceptual overview of MS-Distill. The framework extracts robust spatiotemporal representations by enforcing cross-scale synergistic alignment: (top) temporal invariance via motion continuity across heterogeneous frame rates, and (bottom) structural consistency between holistic skeletons and localized spatiotemporal snippets.

Nonetheless, SSL on structured temporal sequences faces two core challenges. First, contrastive approaches require large batches, memory banks, or complex negative mining [15] to ensure representation diversity; while negatives may be unnecessary [16], avoiding collapse and constructing discriminative embeddings on sparse, highly structured sequences remains difficult. Second, prevailing architectures, such as CNNs or GCNs, impose strong locality biases, restricting their capacity to capture long-range temporal dependencies and multi-scale semantics, e.g., hierarchical compositions from subtle gestures to full-body actions, fundamentally limiting complex behavior understanding.

To address these challenges, we propose **M**ulti-Scale **S**elf-**Distill**ed Learning (**MS-Distill**), a self-supervised framework that models multi-scale spatiotemporal semantics without negative samples, as conceptually illustrated in Fig. 1. MS-Distill adopts a student–teacher distillation paradigm [16] and introduces a multi-view consistency objective to align representations across temporal sampling rates and spatial

*Corresponding author.

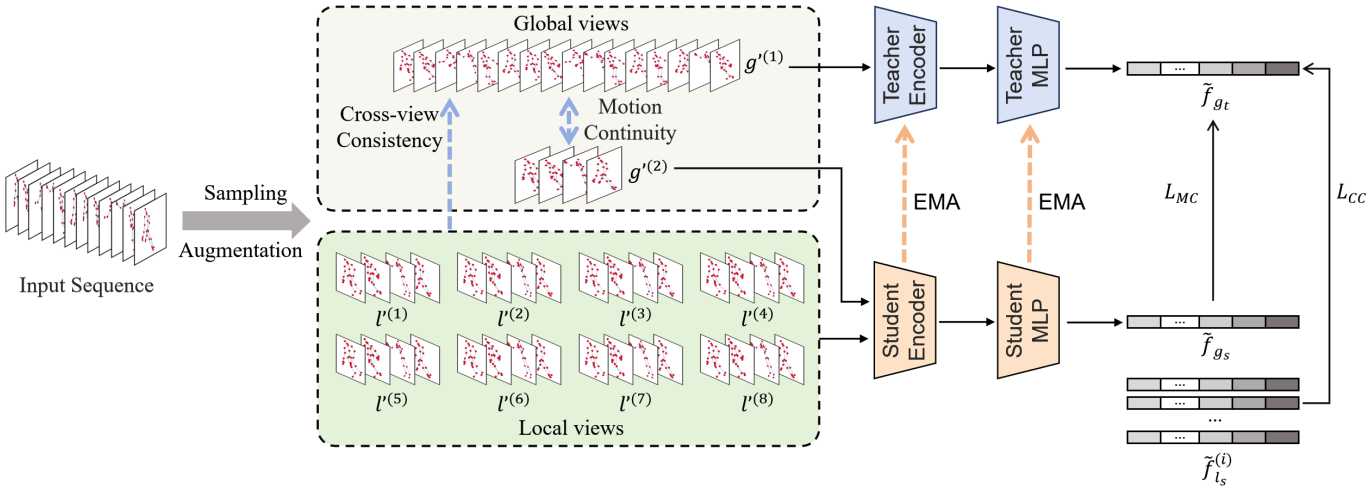

Fig. 2. Overview of MS-Distill. Given an input skeleton sequence, sampling and augmentation generate global and local views. Motion continuity is enforced between two global views, and cross-view consistency is modeled from local to global views. A teacher–student framework with the same Transformer-based encoder and MLP head learns these relationships, where the teacher is updated as the exponential moving average (EMA) of the student.

granularities, enabling the model to coherently capture both fine-grained motions and global action patterns. At its core, a globally attentive spatiotemporal encoder leverages cascaded self-attention modules to explicitly model dependencies between arbitrary spatiotemporal points, overcoming the limitations of local operators and enabling robust hierarchical representation learning across scales.

Our contributions are threefold: (1) We propose **MS-Distill**, a self-supervised multi-scale learning framework that captures both fine-grained motion dynamics and holistic action patterns without relying on negative samples, establishing a unified paradigm for spatiotemporal representation learning. (2) We devise a student–teacher distillation paradigm with a multi-view consistency objective, aligning representations across temporal and spatial hierarchies to produce highly discriminative and broadly generalizable embeddings. (3) Extensive experiments on NTU-RGB+D 60 and 120 demonstrate the superior effectiveness of MS-Distill, e.g., improving NTU-RGB+D 120 accuracy under the KNN evaluation protocol by 3.8% on xset with minimal pre-training, while showing strong transferability to downstream tasks.

The experiments are conducted on publicly available datasets (NTU-RGB+D 60 and 120); however, the source code cannot be released because it depends on proprietary internal libraries that are not publicly accessible.

II. METHOD

3D skeleton sequences play an important role in action recognition, but unsupervised skeleton representation learning remains underexplored. Inspired by the recent method DINO [16], we design MS-Distill, a negative-free self-distillation framework for self-supervised skeleton action recognition. As illustrated in Fig. 2, our spatiotemporal sampling generates global and local views from input skeleton sequences through sampling and augmentation. Global views

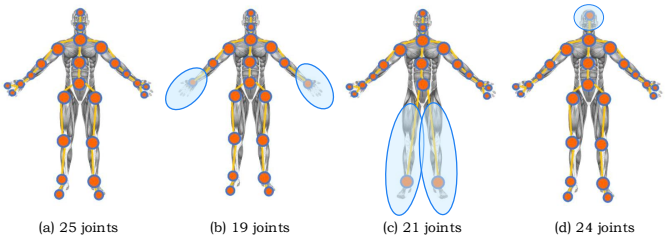

Fig. 3. Schematic diagram of joint sampling in the skeleton sequence space dimension.

are constructed with different frame rates and all joints, while local views are obtained from shorter temporal windows and partial joint subsets. Motion continuity is modeled between two global views, and cross-view consistency is enforced from local to global views.

A randomly selected global view is passed to the teacher encoder to produce target features, while the student encoder processes the other global view together with all local views. Both networks adopt the same Transformer-based encoder followed by an MLP head: the teacher outputs target features, and the student predicts them from online features via a similarity function. The student features are aligned with the teacher using the motion continuity loss (L_{MC}) and the cross-view consistency loss (L_{CC}). Meanwhile, the teacher parameters are updated as the exponential moving average (EMA) of the student's weights, ensuring a stable training process. This design enables MS-Distill to learn robust and generalizable spatiotemporal representations without relying on negative samples.

A. Multi-view skeleton sequence sampling

Given a skeleton sequence $x = \{x_1, \ldots, x_T\}$ with $x_i \in \mathbb{R}^{M\times V\times 3}$ (T frames; M individuals, V body joints per frame,

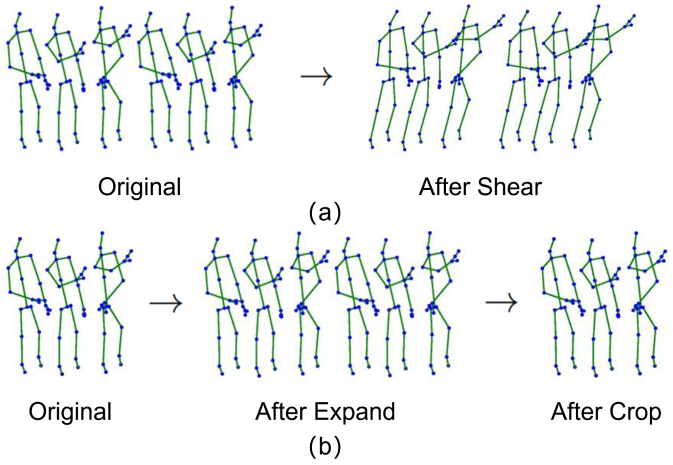

Fig. 4. "Shear" and "Crop" data augmentation for skeleton sequences.

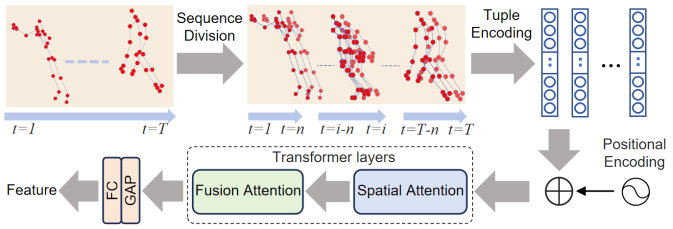

Fig. 5. Schematic diagram of the encoder structure based on Self-Attention, including tuple encoding, position encoding, and spatiotemporal Transformer modules.

and the last dimension representing 3D spatial coordinates), we construct two global and eight local views with different temporal rates and joint subsets. A global view uniformly samples 80% of the timeline at two frame rates ($T \in \{30, 60\}$) while keeping all 25 joints, yielding two global spatiotemporal views $(g^{(1)}, g^{(2)})$. A local view is generated by uniformly sampling frames from a randomly selected region covering $\frac{1}{8}$ of the time axis and by selecting only a subset of the 25 joints (Fig. 3). Using four different frame rates $T \in \{24, 36, 48, 60\}$ yields eight local spatiotemporal views $(l^{(1)}, \ldots, l^{(8)})$. These multi-view clips are used in the student–teacher framework to learn motion continuity between global views and cross-view consistency between local and global views.

B. Data Augmentation

Contrastive learning relies on the shared pattern information [17] between different augmentations of the same sample. To capture such "pattern invariance" in skeleton sequences, we introduce perturbations on the ten sampled skeleton view sequences (two global and eight local), as illustrated in Fig. 4.

Shear [18]: Shear is a linear transformation in the spatial dimension with matrix

$$\mathbf{A} = \begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}, \tag{1}$$

where each a_{ij} is sampled from $[-\beta, \beta]$, and β represents the amplitude of the jitter. Multiplying skeleton coordinates by $\mathbf{A}$ produces randomly tilted postures in the 3D coordinate system (see Fig. 4(a)).

Crop [15]: Cropping perturbs the temporal dimension by asymmetrically inserting frames into the sequence and then randomly deleting some frames to keep the overall length unchanged. This allows the model to adapt to temporal variations and improves its ability to capture action details (see Fig. 4(b)).

After data enhancement, the multi-view skeleton sequence $g'^{(1)}, g'^{(2)}, l'^{(1)}, \ldots, l'^{(8)}$ is fed into the teacher–student framework of MS-Distill to learn spatiotemporal features across views and enhance self-supervised representation learning.

C. Encoding of backbone sequences

To effectively model the long-term dependency of skeleton sequences and the global correlation of spatiotemporal joints, we design a Transformer-based spatiotemporal encoder with tuple and positional encodings. The detailed computational steps of the encoder are summarized in Algorithm 1, while the overall processing pipeline is illustrated in Fig. 5. Specifically, the skeleton sequence of a given action category is first divided into several subsequences (sub-actions) by frame, and each sub-action contains n ($n = 6$) continuous frames. We then use the tuple encoding module to encode each sub-action. Similar to ViT [19], positional embeddings are added using sine and cosine functions of different frequencies to encode and mark each joint, avoiding the loss of joint order information due to tuple encoding:

$$\begin{aligned} PE(p, 2i) &= \sin(p/10000^{2i/C_{in}}) \\ PE(p, 2i+1) &= \cos(p/10000^{2i/C_{in}}) \end{aligned} \tag{2}$$

Here p denotes the joint index and i the embedding dimension. Encoded subsequences are then fed into stacked Transformer layers with spatial attention (SA) to model intra-frame joint relations and fusion temporal attention (FA) to capture inter-frame dynamics. Finally, a global average pooling and a linear layer output the final spatiotemporal representation.

As mentioned, our capture of motion continuity and cross-view consistency involves skeleton views with different spatial

Algorithm 1 Transformer-based encoder module.

Require: Encoded sequence $X_{in} \in \mathbb{R}^{C \times T \times V}$;
Ensure: Features after processing;
1: Compute the Attention:
$Q, K, V = Conv_{2D(1\times1)}(X_{in})$;
$X_{attn} = Tanh(\frac{QK^T}{\sqrt{C}})V$;
2: Concatenate the multiple groups of Attention:
$X_{Attn} = Concat(X^1_{attn}, ..., X^h_{attn})$;
3: Compute the Spatial Attention:
$X_{SA} = Conv_{2D(1\times k_1)}(X_{Attn})$
4: Compute the Fusion Temporal Attention:
$X_{FA} = Conv_{2D(k_2\times1)}(X_{SA})$;
5: **return** X_{out};

and temporal resolutions. We fix the encoded skeleton spatiotemporal vectors to the highest resolution in each dimension and use "interpolation" to handle views of different resolutions during model training to account for the lost spatiotemporal information at lower frame rates or spatial sizes. This allows us to handle inputs of different resolutions using only a single network structure while also embedding the entire model dynamic and making it more suitable for inputs of various sizes in downstream tasks.

D. Motion continuity and cross-view consistency

Our motivation for predicting different views of a skeleton sequence is to let the model learn contextual information by modeling motion continuity (global-to-global) and cross-view consistency (local-to-global). This enables the model to capture the changes in action on the timeline and the consistency between viewpoint transformations, thereby learning motion-invariant and viewpoint-invariant features. Through self-distillation, the model narrows the feature gap between spatiotemporal views and enhances generalization.

Motion Continuity (MC): Similar to video data, the frame rate of skeleton sequences changes motion context (e.g., "walking slowly" vs. "walking fast") and reveals subtle movements. Given two clips with different frame rates, predicting one from the other in feature space compels the model to bridge temporal gaps, thereby effectively capturing long-range temporal dependencies and establishing relations from low- to high-frame-rate inputs. We design a similarity loss by matching the normalized global features $\tilde{f}_{g_t}$ from the teacher and $\tilde{f}_{g_s}$ from the student:

$$\mathcal{L}_{MC} = -\tilde{f}_{g_t} \cdot \log(\tilde{f}_{g_s}), \tag{3}$$

where $\tilde{f}_{g_t}$ and $\tilde{f}_{g_s}$ are normalized feature vectors, and $[\cdot]$ denotes the dot product.

Cross-view Consistency (CC): Besides motion continuity, we also model invariance under multiple views by enforcing local-to-global prediction, which encourages higher-level context learning, including spatial context (masked joints) and temporal context (past or future frames). The loss matches the local features $[\tilde{f}_{l_s}^{(1)}, \ldots, \tilde{f}_{l_s}^{(8)}]$ from the student with the global feature $\tilde{f}_{g_t}$ from the teacher:

$$\mathcal{L}_{CC} = \sum_{i=1}^{k} -\tilde{f}_{g_t} \cdot \log(\tilde{f}_{l_s}^{(i)}), \tag{4}$$

with $k = 8$ local spatiotemporal views.

The final objective combines both terms:

$$\mathcal{L} = \mathcal{L}_{MC} + \mathcal{L}_{CC}. \tag{5}$$

E. Self-supervised training

We train MS-Distill in a self-supervised manner by passing different spatiotemporal views through the student–teacher networks and projection layers to obtain high-dimensional features and predict each other in the latent space. In each training stage, the student parameters θ are updated by back-propagation, while the teacher parameters ξ are updated as the exponential moving average (EMA) of the student [16]:

$$\xi_{t+1} \leftarrow \tau\xi_t + (1-\tau)\theta_t, \tag{6}$$

where $\tau \in [0, 1]$ controls the update speed.

To avoid collapse, we adopt Sharpening and Centering [16]. Sharpening uses a small temperature τ_t to normalize teacher outputs, highlighting highly correlated features and reducing the effect of noise:

$$\tilde{f}_t[i] = \frac{\exp(f_t[i]/\tau_t)}{\sum_{j=1}^{n} \exp(f_t[j]/\tau_t)}, \tag{7}$$

where n is the feature dimension, while Centering adds a bias term c to the teacher output, $f_t \leftarrow f_t + c$, where c is updated by an exponential moving average over the batch mean:

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} f_t[i], \tag{8}$$

where $m > 0$ is the update rate and B the batch size. These operations prevent trivial constant outputs and stabilize training.

III. Experiments and Discussion

A. Datasets

NTU-RGB+D 60 Dataset [20]: This large-scale dataset contains 56,000 action clips across 60 classes, performed by 40 volunteers under three camera views. Each sequence provides 3D coordinates (X, Y, Z) for 25 body joints per subject, with at most two subjects per clip. We strictly follow the two standard evaluation protocols: cross-subject (xsub) and cross-view (xview).

NTU-RGB+D 120 Dataset [21]: As an extension of NTU-60, it scales up to 120 action classes and 113,945 sequences. We evaluate our method under its two standard protocols: cross-subject (xsub) and cross-setup (xset).

B. Experimental setup

Pre-training setup. We pre-train MS-Distill without any labels on the training set of the respective dataset (NTU-60 or NTU-120). For a fair comparison with SkeletonCLR [15], we use the same data pre-processing and a batch size of 128. We randomly initialize the weights associated with spatiotemporal attention and follow the initialization settings in [16]. We use the Adam optimizer [22] with a learning rate set to 5×10^{-4}. We also employ a cosine schedule for learning rate scaling and a weight decay strategy from 0.04 to 0.1. Our work is trained on 2 NVIDIA RTX 1080 Ti GPUs using PyTorch [23].

Downstream task evaluation. After the upstream pre-training is completed, we assess the quality of the learned representations through three standard evaluation protocols:

1) KNN evaluation protocol: This protocol utilizes a K-Nearest Neighbor (KNN) classifier to perform classification based on feature similarity without learning additional weights. We perform a nearest neighbor search on the features generated by the pre-trained encoder. This unsupervised evaluation

TABLE I
RESULTS OF KNN EVALUATION ON NTU-60 AND NTU-120 DATASETS COMPARED WITH THE BASELINE SKELETONCLR

Method	Neg-Free Support	NTU-60 (%)		NTU-120 (%)	
		xsub	xview	xsub	xset
SkeletonCLR [15]	No	57.7	67.3	44.9	45.8
MS-Distill (Ours)	**Yes**	**60.1**	**69.3**	**46.7**	**49.6**

TABLE II
RESULTS OF LINEAR EVALUATION ON NTU-60 AND NTU-120 DATASETS. † INDICATES METHODS USING NEGATIVE SAMPLES.

Method	Neg-Free Support	NTU-60 (%)		NTU-120 (%)	
		xsub	xview	xsub	xset
LongT GAN [11]	Yes	39.1	48.1	35.6	39.7
PCRP [24]†	No	53.9	63.5	41.7	45.1
AS-CAL [18]†	No	58.5	64.8	48.6	49.2
SkeletonCLR [15]†	No	68.3	76.4	56.8	55.9
MG-AL [25]	Yes	64.7	68.0	46.2	49.5
CRRL [26]†	No	67.6	73.8	56.2	57.0
SDS-CL [27]†	No	**73.6**	**78.9**	50.6	55.6
VCL [28]†	No	71.0	76.8	57.5	58.8
MS-Distill (Ours)	**Yes**	69.6	77.5	**57.7**	**60.2**

directly reflects the intrinsic quality and cluster structure of the learned feature space, demonstrating the model's ability to group similar actions together.

2) Linear evaluation protocol: Under this protocol, we freeze the pre-trained encoder and perform downstream task classification based on its generated features. Specifically, a linear classifier (consisting of fully connected and Softmax layers) is trained on top of the frozen features. Since the encoder parameters remain unchanged, this protocol strictly tests the discriminative ability of the static features learned by the encoder during the self-supervised stage.

3) Fine-tuning protocol: In this protocol, we attach a linear classifier to the pre-trained encoder and fine-tune the entire network end-to-end on the labeled dataset. This approach allows the model to adapt its parameters to the specific action recognition task. The performance here reflects the transferability of the pre-trained weights and is typically compared against the linear evaluation results to gauge the benefit of feature adaptation.

C. Comparison with advanced methods

We compare MS-Distill with state-of-the-art methods on NTU-60 and NTU-120 using KNN, linear, and fine-tuning protocols.

KNN evaluation. As shown in Table I, MS-Distill surpasses SkeletonCLR [15] on both datasets with the same k ($k = 20$), indicating more discriminative features.

Linear evaluation. Table II presents the comparisons with state-of-the-art methods. On the smaller NTU-60 dataset, SDS-CL [27] yields higher accuracy, largely benefiting from its multi-stream training strategy that fuses joint and motion modalities. In contrast, MS-Distill operates as a pure single-stream framework, extracting features solely from joint coordinates. Regarding VCL [28], although it achieves competitive

TABLE III
RESULTS OF FINE-TUNING ON NTU-60 AND NTU-120 DATASETS.

Method	NTU-60 (%)		NTU-120 (%)	
	xsub	xview	xsub	xset
ST-GCN [29] (Random Init.)	82.8	91.0	76.4	76.8
MS-Distill + ST-GCN	**83.0**	**91.1**	**76.7**	**77.2**

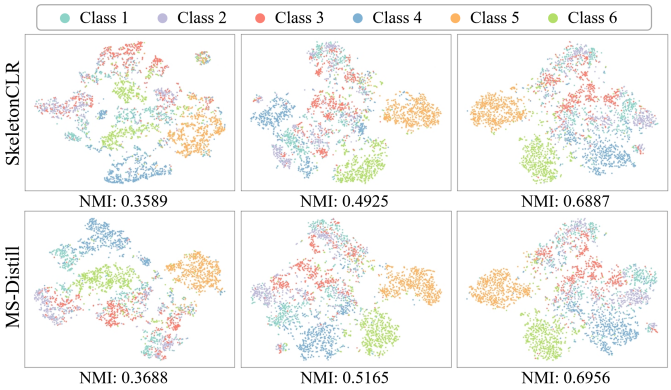

Fig. 6. t-SNE visualization of embeddings for six different action classes. Different colors represent different ground-truth action categories. MS-Distill (bottom) shows significantly tighter clusters and better class separation than SkeletonCLR (top).

results on NTU-60, it fundamentally relies on a contrastive paradigm necessitating a memory bank to maintain negative samples. Notably, MS-Distill outperforms VCL on the NTU-60 xview protocol (77.5% vs. 76.8%), indicating superior robustness to viewpoint variations without the architectural constraint of requiring negative pairs. The superiority of our method is more prominently demonstrated on the large-scale NTU-120 dataset. Since NTU-120 contains greater variation and demands higher generalization capability, these results highlight the scalability of our approach. Specifically, under the challenging cross-setup (xset) protocol, our method surpasses the nearest competitor (VCL) by 1.4% and outperforms SDS-CL by 4.6%, achieving new state-of-the-art performance without the computational and memory overhead of maintaining a large memory bank.

Fine-tuning. As shown in Table III, MS-Distill performs better than supervised ST-GCN under the same structure and parameters, verifying that the framework can learn richer information for better weight initialization.

Qualitative results. We use t-SNE [30] and fixed settings to show the distribution of features obtained after the encoder. We randomly selected 6 action categories and used t-SNE visualization for a fair comparison. As shown in Fig. 6, from the visual results and the calculated normalized mutual information (NMI), we can conclude that our MS-Distill can cluster skeleton feature embeddings with the same category labels more tightly than SkeletonCLR [15]. At the same time, MS-Distill can achieve better performance with fewer pre-training epochs, which shows the efficiency of our model in the pre-training process.

TABLE IV

ABLATION STUDY OF MULTI-VIEW PREDICTION STRATEGIES FOR NTU-60 LINEAR EVALUATION. l AND g DENOTE LOCAL AND GLOBAL VIEWS, RESPECTIVELY. $a \rightarrow b$ SIGNIFIES PREDICTING VIEW b FROM VIEW a.

Strategy Description	Configurations				NTU-60 (%)	
	$l \rightarrow g$	$g \rightarrow g$	$l \rightarrow l$	$g \rightarrow l$	xsub	xview
Baseline (Global only)	✗	✓	✗	✗	67.2	75.8
Local-to-Global (Single)	✓	✗	✗	✗	68.4	77.1
Multi-Scale Distill (Ours)	✓	✓	✗	✗	**69.8**	**77.5**
Cross-Scale (All-in)	✓	✓	✓	✗	68.9	76.8

D. Ablation experiment

Perspective relevance. Table IV compares different view-prediction strategies. Jointly predicting local-to-global and global-to-global correspondences yields the best accuracy, while global-to-local or local-to-local prediction degrades performance, as the model overfits local details and loses contextual information.

IV. CONCLUSION

We present Multi-Scale Self-Distilled Learning (**MS-Distill**), a self-supervised framework that captures hierarchical spatiotemporal dynamics from skeleton sequences. By aligning multi-view representations across temporal scales and joint subsets within a student–teacher architecture, MS-Distill learns robust motion- and viewpoint-invariant features. A globally attentive spatiotemporal encoder explicitly models long-range dependencies, unifying fine-grained motions and holistic action patterns. Extensive evaluations on NTU-RGB+D 60 and 120 demonstrate substantial gains in recognition accuracy. Notably, as mentioned in our analysis, our method is fast to train (converges in just a few iterations). It does not require negative samples and large batches of data needed by previous comparative skeleton representation learning methods, thus maintaining high efficiency.

Limitation. Our method currently only processes the joint data stream of the skeleton sequence data. Although other data streams of the skeleton (e.g., bone and motion information) are readily available and complementary, simple integrated learning fails to effectively model their interactions. In future work, we will investigate how to leverage this complementary information to better fuse data across different streams.

REFERENCES

[1] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "From cnns to transformers in multimodal human action recognition: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 8, Jul. 2024.

[2] X. Wang, B. Lafreniere, and J. Zhao, "Exploring visualizations for precisely guiding bare hand gestures in virtual reality," in *Proc. CHI*, 2024.

[3] M. R. Mishra and P. K. Meher, "Abnormal human behaviour detection by surveillance camera," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 2024, pp. 1–6.

[4] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[5] W. Jia, H. Wang, Q. Chen, T. Bao, and Y. Sun, "Analysis of kinect-based human motion capture accuracy using skeletal cosine similarity metrics," *Sensors*, vol. 25, no. 4, 2025.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. CVPR*, 2017, pp. 7291–7299.

[7] Z. Liao, J. Zhu, C. Wang, H. Hu, and S. L. Waslander, "Multiple view geometry transformers for 3d human pose estimation," in *Proc. CVPR*, June 2024, pp. 708–717.

[8] H. Liu, Y. Li, T.-J. Mu, and S.-M. Hu, "Recovering complete actions for cross-dataset skeleton action recognition," in *Proc. NeurIPS*, vol. 37, 2024, pp. 92 055–92 081.

[9] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "Blockgcn: Redefine topology awareness for skeleton-based action recognition," in *Proc. CVPR*, June 2024, pp. 2049–2058.

[10] J. Do and M. Kim, "Skateformer: Skeletal-temporal transformer for human action recognition," in *Computer Vision – ECCV 2024*, 2025, pp. 401–420.

[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*, vol. 32, no. 1, 2018.

[12] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proc. CVPR*, 2020, pp. 9631–9640.

[13] M. Abdelfattah, M. Hassan, and A. Alahi, "Maskclr: Attention-guided contrastive learning for robust action representation learning," in *Proc. CVPR*, June 2024, pp. 18 678–18 687.

[14] X. Yu, H. Miao, C. Pang, and L. Lyu, "Adaptive koopman contrastive learning for skeleton-based action recognition," *Neurocomputing*, vol. 658, p. 131750, 2025.

[15] 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. CVPR*, 2021, pp. 4741–4750.

[16] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, 2021, pp. 9650–9660.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, 2020, pp. 9729–9738.

[18] 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, pp. 90–109, 2021.

[19] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proc. CVPR*, 2016, pp. 1010–1019.

[21] 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.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Proc. NeurIPS*, vol. 32, 2019.

[24] S. Xu, H. Rao, X. Hu, J. Cheng, and B. Hu, "Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition," *IEEE Transactions on Multimedia*, 2021.

[25] Y. Yang, G. Liu, and X. Gao, "Motion guided attention learning for self-supervised 3d human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8623–8634, 2022.

[26] P. Wang, J. Wen, C. Si, Y. Qian, and L. Wang, "Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 6224–6238, 2022.

[27] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11 035–11 048, 2024.

[28] D. D. Nguyen, D. A. Latif, and T. Zaharia, "Variational contrastive learning for skeleton-based action recognition," 2026. [Online]. Available: https://arxiv.org/abs/2601.07666

[29] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.