

Open Your Eyes to See More: Dual Perspective Contrastive Learning for Skeleton-Based Action Understanding

Mahmoud Ali¹, Di Yang^{1,2}, Quan Kong³, Gianpiero Francesca⁴ and François Brémond¹

¹ Inria Center at Université Côte d’Azur, Valbonne, France

² Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China

³ Woven by Toyota, Tokyo, Japan

⁴ Toyota Motor Europe, Brussels, Belgium

Abstract—Self-supervised learning has emerged as a powerful approach for skeleton-based action recognition, with contrastive methods driving recent progress. However, most existing approaches use a single encoder to jointly model spatial and temporal features, which can blur their distinct semantics and hinder fine-grained motion understanding. To tackle this, we propose SKEYES, a dual-perspective self-supervised framework with two parallel encoders that separately capture spatial and temporal dynamics. This design avoids early feature fusion and preserves the unique characteristics of pose and motion. We further introduce a dual contrastive learning objective that aligns both intra-view and cross-view features, which can promote complementary learning across feature types. To ensure efficiency, only the main encoder is used during inference. Extensive experiments on six benchmark datasets covering both laboratory settings (NTU RGB+D 60/120, PKU-MMD) and real-world environments (Toyota SmartHome, Penn Action, Posetics) demonstrate that SKEYES achieves state-of-the-art performance when transferring for action recognition and action detection tasks, with strong generalization even under low-label conditions.

I. INTRODUCTION

Human action recognition from skeleton data [41], [44], [4], [54], [12], [10] has made promising progress due to the robustness to changes in light conditions, viewpoints, appearances, etc. Recently, self-supervised learning (SSL) has recently achieved great success in skeleton-based action understanding [19], [43], [45], [48], [33], [27], [32], [26], [39]. By learning from unlabeled pose sequences, these methods can capture meaningful patterns in human motion and significantly reduce the need for costly annotations. Among them, contrastive learning [19], [39] has shown strong potential by encouraging the model to pull together positive pairs and push apart negative ones, helping it discover useful representations. These advances have made SSL a promising solution for many applications like action recognition, action retrieval, and even action detection.

However, despite this progress, current methods still face important challenges. A major issue lies in how these models treat spatial and temporal information. Most existing approaches [19], [45], [39] use a single encoder to jointly learn from both the skeleton (spatial) and motion (temporal) signals. Yet, these two types of information are quite different in nature: spatial features describe body configurations at a

moment in time, while temporal features capture how motion evolves over time. Mixing them too early can confuse the learning process and limit how well the model understands complex actions. Another challenge comes from the way contrastive learning is applied. Many methods [19], [39] only compare features of the same type, e.g., spatial features with spatial ones, or global features with global ones. This limits the model ability to learn relationships across different levels of information and prevents it from seeing the bigger picture.

To solve these problems, we propose a new self-supervised framework based on dual-perspective contrastive learning, namely SKEYES. As shown in Fig. 1, inspired by how humans use both eyes to perceive depth and detail from different perspectives, our framework uses two strongly augmented views of the same skeleton sequence, processed by two parallel encoders, defined as a main encoder and an auxiliary encoder. The auxiliary encoder can be either homogeneous or heterogeneous. These encoders extract features separately for spatial and temporal information using dedicated projection modules. This design helps the model better focus on each type of information without mixing them too early, addressing the mismatch between spatial and temporal dimension. During the transfer learning stage for downstream tasks, the auxiliary encoder is discarded, only the main encoder is utilized, thereby preventing any increase in computational complexity and maintaining efficiency during inference.

We further design a dual contrastive learning objective to make the most of the learned features. First, we ensure that the spatial and temporal features extracted from the same view remain consistent with each other, so the model will not learn redundant or conflicting signals. Then, we go a step further and create contrastive pairs between different types of features across the two views: for example, comparing spatial features from one view with global features from the other. This dual-perspective alignment helps the model discover complementary patterns that would not be visible with traditional contrastive learning. In addition, we build global representations by combining spatial and temporal cues, which further strengthen the learned features for downstream tasks.

We evaluate our approach on several challenging benchmarks, including both laboratory scenes, e.g., NTU RGB+D 60 [28] and NTU RGB+D 120 [25], PKU-MMD [8], and real-world scenarios, e.g., Posetics [44], Toyota SmartHome [9] and Penn Action [52]. These datasets cover a wide range

This work was supported by Toyota Motor Europe (TME) and the French government, through the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002; and performed using HPC resources from GENCI-IDRIS (Grant 2026-AD011015057R2).

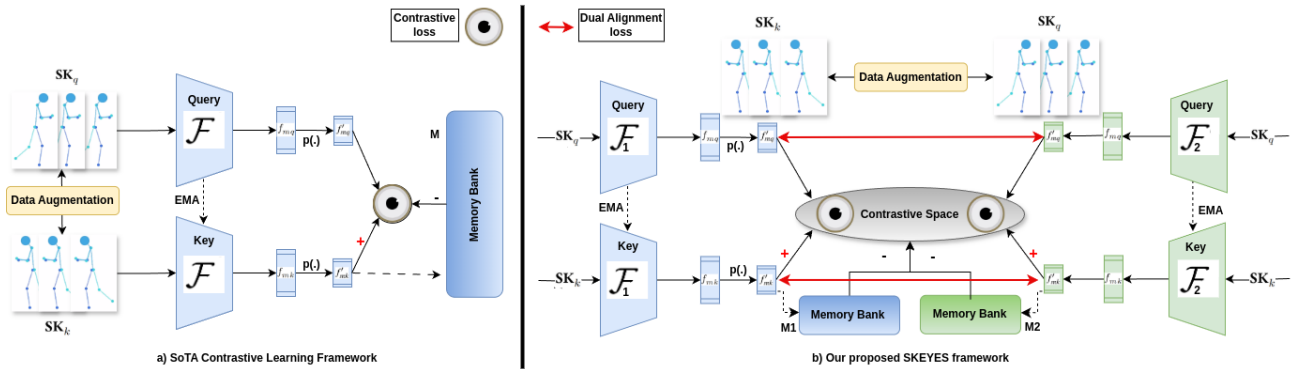


Fig. 1. Framework for our proposed SSL skeleton-based action representation learning.

of human actions and environments. Following standard self-supervised protocols, we test our model on tasks such as action recognition and transfer learning. Our method outperforms existing SSL models and achieves state-of-the-art performance,

In summary, our main contributions are as follows: (i) we introduce SKEYES, a novel skeleton-based SSL framework with dual encoders. (ii) We propose a novel dual-perspective contrastive learning strategy to handle spatial-temporal mismatches. We design a dual-perspective alignment to enhance feature complementarity. (iii) We conduct extensive experiments and analysis to validate the effectiveness of our method across a good number of datasets and tasks, including action classification and action detection.

II. RELATED WORK

a) Spatio-Temporal Skeleton Encoders: Recognizing the natural graph-like structure of skeletons, ST-GCN [41] and its successors [16], [29], [41], [1], [4] innovatively applied Graph Neural Networks (GCNs), crafting convolution kernels tailored to the skeleton topology learning. This groundbreaking framework catalyzed a wave of GCN-driven advancements in skeleton action recognition. To improve flexibility and adapt to real-world conditions such as occlusion, missing joints, or unseen skeleton formats, topology-free models have recently been proposed. These approaches [44], [12], [54], [10] leverage Transformers [54], [10] or Convolutional Networks [44], [12] to model sequences without relying on predefined joint connections and show strong generalization capabilities in scenarios with diverse viewpoints and compositional actions. However, the mentioned methods still suffer from two major limitations: (1) the lack of pretraining hampers generalization to real-world conditions involving viewpoint shifts and compositional actions, and (2) the global feature extraction architectures struggle to capture fine-grained spatial and temporal dependencies.

SKEYES addresses both challenges by introducing a dual-perspective self-supervised framework with separate spatial and temporal encoders. This disentangled design preserves modality-specific features and avoids early fusion. SKEYES shows strong performance across both controlled and in-the-wild datasets.

b) Self-Supervised Skeleton Representation: Skeleton-based self-supervised learning typically follows a two-stage paradigm: pretraining with a pretext task, followed by

fine-tuning on downstream applications. Among existing approaches, generative approaches focus on generating [55], [15] or reconstructing the masked [26] input data, while contrastive learning has become dominant, to learn discriminative features by pulling positive pairs together and pushing negative pairs apart in a latent space [19], [45], [27], [39]. Although these techniques have demonstrated impressive results on curated 3D datasets, they exhibit significant limitations when applied to in-the-wild 2D skeletons. Their performance typically degrades due to noisy joint positions, occlusions, and viewpoint changes [9], [21], [52]. Furthermore, most contrastive schemes treat all input frames or modalities equally, without accounting for the varying informativeness of cues across space-time.

SKEYES addresses these issues by introducing a dual-perspective self-supervised framework that disentangles spatial and temporal modeling using two separate encoders. Then, the dual contrastive learning objective further aligns intra-view and cross-view representations, allowing the model to learn complementary and robust features. SKEYES selectively preserves and aligns complementary representations, to learn richer and more transferable skeleton representations, especially under challenging real-world conditions.

III. PROPOSED APPROACH

In this section, we present our proposed SKEYES, a dual-perspective self-supervised representation learning framework that improves skeleton-based action recognition.

Overview: As shown in Fig. 2, the proposed framework, SKEYES, includes three key components. (1) Skeleton Sequence Modeling to generate diverse views of input skeleton sequences through a set of data augmentation strategies. (2) Dual Perspective Skeleton Representation to extract skeleton representations and decouple into separate spatial and temporal components. These features are then refined using a reasoning model. (3) Dual Alignment Contrastive Loss, where an Inter Dual Alignment Loss (DAL) enforces consistency between the dual branches, and an Intra Contrastive Loss (CoL) enhances feature discrimination through a memory bank within each branch. In the following subsection, we detail the three main components of our framework.

A. Skeleton Sequence Modeling

Given a skeleton sequence $\mathbf{sk} \in \mathbb{R}^{T \times J \times C_{in}}$, where T is the sequence length, J is the number of body joints per frame, and

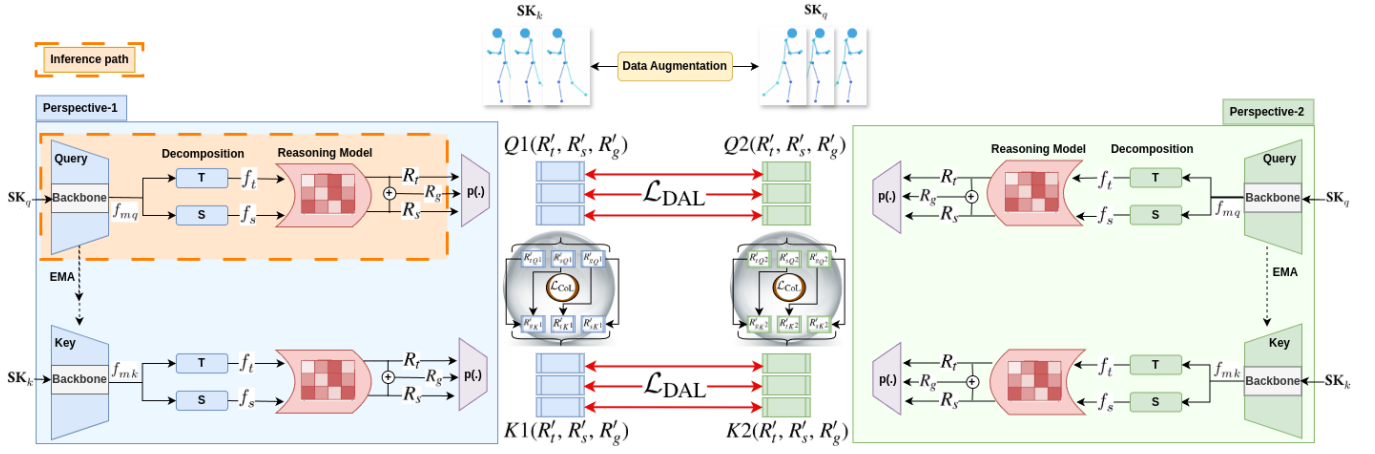


Fig. 2. Overview of SKEYES framework. Given a skeleton sequence SK , we generate two augmented views, SK_q and SK_k . These are processed using the dual-parallel branches (left and right) shown in this Figure. In each branch, the Query and Key are functionally identical. First, motion features f_m are extracted from SK_q . Second, these features are decoupled into temporal (f_t) and spatial (f_s) components. Third, both components are refined using a reasoning model to obtain R_t and R_s , which are then fused into a global representation R_g . Fourth, all representations (R_t , R_s , and R_g) are projected into a lower-dimensional space, resulting in R'_t , R'_s , and R'_g . Finally, the dual-branch query and key representations are $(Q_1(R'_t, R'_s, R'_g))$, $(K_1(R'_t, R'_s, R'_g))$ and (Q_2, K_2) respectively. These representations are optimized using two complementary objectives: *Inter-Dual Alignment Loss (DAL)*, which aligns features across the two branches, and *Intra-Contrastive Loss (CoL)*, which encourages discriminative learning within each branch.

C_{in} are the joint coordinates (either 2D or 3D). We adopt the data augmentation strategy to generate two different views of the same skeleton sequence, denoted as a query sequence SK_q and a key sequence SK_k , respectively. Following [39], this data augmentation process includes standard transformations such as rotation, flipping, and spatial and temporal masking to enhance the robustness of the features.

B. Dual Perspective Skeleton Representation

Unlike existing methods, our proposed framework **SKEYES** employs a *dual-parallel branch architecture* to process skeleton sequences, as illustrated in Figure 2. The augmented features SK_q and SK_k are fed into a dual-branch framework. Each branch is based on the Momentum Contrast MoCo-v2 [3] paradigm, where both the query and key encoders share identical architectures, but are updated differently to maintain consistency. The query encoder is trained using standard backpropagation, while the key encoder is updated via a momentum-based moving average of the query encoder’s parameters. Each pipeline is composed of three main modules: a *motion backbone* that captures the dynamic patterns in the skeleton, a *feature decomposition module* that disentangles the motion features into spatial and temporal components, and a *feature refinement module* that enhances these components using reasoning mechanisms such as Transformers. **SKEYES** is designed to be modular and plug-and-play, allowing the backbone encoder to be easily replaced with any state-of-the-art architecture. This flexibility enables compatibility with various GCN-based, graph-free, or Transformer-based models, facilitating adaptation to diverse datasets and computational constraints. These dual encoders can be homogenized, sharing the same architecture but initialized differently, or un-homogenized, using distinct encoder architectures. This dual setup serves as a form of data augmentation—not at the skeleton sequence level, but in the latent space—enhancing representation diversity and

enabling the model to capture more information from the skeleton sequence from multiple perspectives.

Specifically, **Motion backbone**: Given an input sequence SK_q , we first extract intermediate features using a backbone network. This backbone can be selected from various state-of-the-art models, including GCN-based architectures [5], [7], [18], graph-free approaches like [44], or Transformer-based models such as [53], [10]. The backbone encodes the spatio-temporal dynamics of the skeleton sequence, capturing both joint configurations and motion trajectories. **Feature Decomposition**: As shown in prior studies [39], [11] that directly extracting spatio-temporal features from encoders contains entangled representations that are suboptimal for contrastive learning. We address this by decoupling the features into separate temporal (f_t) and spatial (f_s) components through MLP layers (Eq. 1), resulting in more discriminative representations for action recognition. **Feature Refinement**: To better capture complex motion dynamics in different scenarios like real-world activities involving both short- and long-duration actions, we incorporate a reasoning module applied to spatial (f_s) and temporal (f_t) features. This Transformer-based module captures both localized short-term dependencies and long-range interactions through dense self-attention, producing refined representations R_s and R_t by applying Eq. 2.

We fill the gap between the temporal and spatial features by leveraging both features together using the concatenation of each model’s respective features, producing global representations $R_g = \text{Concat}(R_t, R_s)$.

These three features (R_t , R_s , R_g) are then projected into a lower-dimensional space suitable for contrastive learning via a projection function $p(\cdot)$, resulting in $Q_{R'_t}, Q_{R'_s}, Q_{R'_g}$ for Transformer based features. The same process is repeated for the key branch to produce $K_{R'_t}, K_{R'_s}, K_{R'_g}$.

This architectural design is central to our proposed **SKEYES** framework and enables the subsequent dual align-

ment contrastive learning strategies described next.

$$f_t, f_s = \text{ReLU}(\text{LayerNorm}(\text{Linear}(fm))) \quad (1)$$

$$\begin{aligned} Q_T = & \text{LN}(X + \text{MHA}(Q, K, V)) \\ & + \text{FFN}\left(\text{LN}(X + \text{MHA}(Q, K, V))\right) \end{aligned} \quad (2)$$

C. Dual Alignment Contrastive Loss

To fully exploit the complementary strengths of our dual-perspective architecture, we introduce a *two Objective learning module* that refines multi-granularity features through Inter Dual alignment and Intra contrastive separation.

Objective 1 - Inter Dual Alignment: To leverage the full potential of our dual-perspective models, it is crucial to enhance the complementary features by reducing the conflicting, redundant, and noisy cues. We enforce this by maintaining the consistency between both branches’ representations using *Inter Dual Alignment Loss (DAL)*. DAL minimizes the N -dimensional embedding dissimilarities between the corresponding pairs from two branch features from the query and key in each branch: $(R1'_t, R2'_t)$, $(R1'_s, R2'_s)$, and $(R1'_g, R2'_g)$ by applying standard mean squared error loss as in Eq. 3.

$$\mathcal{L}_{\text{Inter-DAL}} = \frac{1}{N} \sum_{i=1}^N (R1'_i - R2'_i)^2 \quad (3)$$

This process not only improves overall feature quality but also indirectly improves contrastive learning by improving the quality of positive and negative sample pairs. As a result, both positive and negative samples benefit from enriched feature representations, leading to more effective contrastive learning.

Objective 2 - Intra Contrastive Learning: We follow MoCo-v2 [3] for contrastive learning to enhance the representations on the branch intraly by bringing similar features (*positive pairs*) closer by minimizing their distance while pushing apart dissimilar features (*negative pairs*) stored in the memory bank. In contrastive space, we aim to project the most meaningful embedding obtained from SKEYES to ensure superior separation among positive and negative pairs. ICoL has a critical role in enriching the embedding space by augmenting the All-pair reasoning. The final representations from both the *query* and *key* in the branch utilized by InfoNCE contrastive loss, which effectively enhances feature discrimination.

Given that we have different feature representations—**spatial (s), temporal (t), and global (g)**—we follow [39] and apply **Intra-CoL (Intra Contrastive Learning)** Eq 4 in mixed feature pairs within the branch rather than in instance feature pairs, as this has shown good effectiveness. Our contrastive loss is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{Intra-CoL}} = & \mathcal{L}_{\text{info}}(Q_s, K_g) + \mathcal{L}_{\text{info}}(Q_t, K_g) \\ & + \mathcal{L}_{\text{info}}(Q_g, K_s) + \mathcal{L}_{\text{info}}(Q_g, K_t) \end{aligned} \quad (4)$$

where $\mathcal{L}_{\text{info}}$ represents the InfoNCE loss, and Q and K denote the features from the query and key pipelines, respectively. The InfoNCE loss is defined as:

$$\mathcal{L}_{\text{info}}(Q, K) = -\log \frac{\exp(Q \cdot K^+ / \tau)}{\sum_{K^-} \exp(Q \cdot K^- / \tau)} \quad (5)$$

where $Q \cdot K^+$ represents the similarity between the query and the positive keys, K^- represents negative samples from the Transformer memory bank, τ is a temperature scaling parameter that controls the sharpness of the similarity distribution.

$$\mathcal{L} = \mathcal{L}_{\text{Inter-DAL}} + \mathcal{L}_{\text{Intra-CoL}} \quad (6)$$

Finally, our self-supervised pretraining loss function $\mathcal{L}_{\text{DACO}}$ combines both objectives and is formulated as shown in Eq 6. This dual-objective strategy ensures that the learned representations are both internally consistent and externally discriminative, forming a strong foundation for downstream tasks.

IV. EXPERIMENTS

We conduct extensive experiments to evaluate the effectiveness of SKEYES under different evaluation protocols, as shown in Section IV-B. Firstly, we compare SKEYES with state-of-the-art (SoTA) self-supervised models through various evaluation protocols, including linear evaluation (linear probing), k-NN, semi-supervised learning, and cross-dataset transfer learning, on 2D real-world datasets (*i.e.*, **Posetics** [44], **Toyota SmartHome** [9] (SmartHome) and **Penn Action** [52] and also on 3D datasets (*i.e.*, **NTU-RGB+D 60** [28], **NTU-RGB+D 120** [25], **PKU-MMD II** [8]). Subsequently, we provide deeper insights by conducting per-class analysis. Finally, we present a comprehensive ablation study to highlight the impact of each component within the SKEYES framework. Further studies and experiments, can be found in **the Appendix**.

A. Datasets and Evaluation Metrics

a) **NTU-RGB+D 60** [28]: 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.

b) **NTU-RGB+D 120** [25]: 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.

c) **PKU-MMD** [8]: is an untrimmed laboratory dataset comprising 51 actions, divided into two subsets: Part I and Part II. For action classification on PKU-MMD Part II, action instances are segmented using temporal annotations and split into training and testing sets following the cross-subject evaluation protocol. In our experiments, we use PKU-MMD Part II with 5,332 samples for training and 1,613 for testing. For action detection, we use the untrimmed videos from PKU-MMD Part I and report event-based mean Average Precision (mAP) under different IoU thresholds.

Methods	Posetics	
	Top-1(%)	Top-5(%)
Pr-ViPE [30] (ECCV 20)	17.2	35.3
OR-VPE [46](FG 21)	14.6	31.2
3s-CrosSCLR [20] (CVPR 21)	18.8	38.1
AimCLR [33] (AAAI 22)	19.2	39.3
CMD [27] (ECCV 22)	20.4	40.5
HiCLR [6] (AAAI 23)	20.1	39.9
HiCo [11](AAAI 23)	21.3	42.1
PCM ³ [50](ACMMM 23)	20.0	40.3
ViA [45](IJCV 24)	20.7	40.1
USDRL [38](AAAI 25)	25.9	48.6
SKEYES (Ours)	27.2	50.0

TABLE I

ACTION CLASSIFICATION (LINEAR EVALUATION) ON POSETICS.

Methods	SmartHome		Penn Action
	CS(%)	CV2(%)	Top-1(%)
OR-VPE [46]	42.7	32.4	78.5
AimCLR [33]	46.6	48.3	-
CMD [27]	49.0	52.5	-
HiCLR [6]	49.1	52.3	88.7
ViA [45]	49.5	52.6	90.2
HiCo [11]	54.3	54.8	87.6
PCM ³ [50]	45.3	46.8	85.6
USDRL [38]	54.3	53.9	89.2
SKEYES (Ours)	54.6	56.9	91.0

TABLE II

ACTION CLASSIFICATION (TRANSFER LEARNING EVALUATION "LINEAR") ON SMARTHOME AND PENN ACTION.

d) Posetics [44]: 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.

e) Toyota Smarthome [9]: (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 [42]. 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.

f) Penn Action [52]: 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.

B. Evaluation Protocols

We evaluate the learned representations using six protocols. Firstly, **Linear Evaluation**, a linear action classifier is trained on top of the frozen encoder to predict action labels. **Transfer Learning Evaluation** examines model generalization: on lab datasets, we fine-tune the full model with labeled data as in prior work [19], [39], while on real-world datasets,

Method	NTU-RGB+D 60		NTU-RGB+D 120	
	CS (%)	CV (%)	CS (%)	CSet (%)
ActCLR [22] (CVPR 23)	80.9	86.7	69.0	70.5
HiCLR [6](AAAI 23)	80.4	85.5	70.0	70.4
PCM ³ [50] (ACMMM 23)	83.9	90.4	76.5	77.5
HiCo [11] (AAAI 23)	81.1	88.6	72.8	74.1
Eq-Contrast [23] (TIP 24)	83.9	90.3	75.7	77.2
ViA [45] (IJCV 24)	78.1	85.8	69.2	66.9
ACA2Net [2](TCSVT 25)	86.0	89.6	-	-
ActCLR+ [24](TPAMI 25)	82.3	88.2	70.9	73.2
MaskSem [37](IROS 25)	85.9	90.8	77.5	79.3
Heterogeneous [34] (CVPR 25)	80.2	88.0	70.7	73.5
USDRL [38] (AAAI 25)	84.2	90.8	76.0	76.9
PCM ³ ++ [51] (IJCV 26)	84.8	91.0	76.7	-
SKEYES (Ours)	87.2	92.3	79.1	80.1

TABLE III

ACTION CLASSIFICATION (LINEAR EVALUATION) ON NTU-RGB+D 60 AND 120 DATASETS.

Method	Transfer to PKU-MMD II
	NTU-RGB+D 120 CS(%)
ISC [32]	52.3
HiCo-Transformer [11]	55.4
UmURL-3 [31]	58.5
A ² MC [40]	58.9
SCD-Net [39]	64.0*
Heterogeneous [34]	63.1
SKEYES (Ours)	64.9

TABLE IV

ACTION CLASSIFICATION (TRANSFER LEARNING EVALUATION "FINE-TUNE") ON PKU-MMD II. * MEANS OUR IMPLEMENTATIONS.

only the classifier is retrained after large-scale pretraining following [45]. In **Semi-supervised Evaluation**, the encoder is first pre-trained with unlabeled data and then fine-tuned using 5% and 10% of the labeled data, following standard practice. In **Action retrieval- KNN Evaluation**, a non-parametric K-Nearest Neighbor classifier is used to assess the quality of the representation space. **Action Prediction Evaluation** assessing the model's ability to recognize an action as early as possible by observing only a portion of the skeleton sequence (70%, 80%, or 90%) under the same setting of linear evaluation. Besides the above-mentioned classification tasks for trimmed video datasets, we also perform **Action Detection Evaluation** on an untrimmed video dataset: PKU-MMD Part I. We freeze the encoder, train a linear classifier, and compute mean Average Precision (mAP) across multiple temporal IoU thresholds.

C. Comparisons with State-of-the-Art Methods

a) Action Classification (Linear Evaluation): We report the linear evaluation results of our method, SKEYES, in Tables III and I. Using only the joint modality, SKEYES consistently outperforms previous state-of-the-art methods on both lab-controlled datasets (NTU-RGB+D 60 and NTU-RGB+D 120) and the real-world dataset (Posetics), achieving notable gains across all evaluation protocols. These improvements indicate that SKEYES learns highly discriminative

Methods	SmartHome		Penn Action	
	(5%)	(10%)	(5%)	(10%)
HiCo [11]	34.5	46.0	57.2	74.5
PCM ³ [50]	23.1	30.1	46.3	60.9
ViA [45]	38.6	45.3	65.8	85.2
USDRL [38]	37.1	43.6	58.1	70.3
SKEYES (Ours)	39.0	48.0	72.9	86.5

TABLE V

ACTION CLASSIFICATION (SEMI-SUPERVISED EVALUATION 5% AND 10%) ON SMARTHOME CS(%) AND PENN ACTION.

Method	CS(%)		CV(%)	
	5%	10%	5%	10%
HiCLR [6]	63.3	70.7	68.3	74.8
HiCo [11]	54.4	73.0	54.8	78.3
PCM ³ [50]	-	77.1	-	82.8
A ² MC [40]	-	76.4	-	81.5
UmURL [31]	72.5	-	76.8	-
SCD-Net [39]	79.2	82.2	81.6	85.8
ActCLR+ [24]	-	82.2	-	86.0
SKEYES (Ours)	81.0	83.5	83.4	86.9

TABLE VI

ACTION CLASSIFICATION (SEMI-SUPERVISED EVALUATION 5% AND 10%) ON NTU-RGB+D 60.

and transferable representations, as evidenced by its strong performance under the linear probing setting. Moreover, the consistent gains across diverse benchmarks demonstrate the robustness and generalization capability of SKEYES in both controlled and in-the-wild scenarios.

b) Action Classification (Transfer Learning Evaluation):

We investigate the transferability and generalization capabilities of the representations learned by SKEYES across both real-world and lab-controlled settings using linear and Fine-tune transfer learning evaluation, respectively. For real-world actions, as shown in Table II, we report Top-1 accuracy on Penn Action and use mean per-class accuracy for SmartHome due to class imbalance. SKEYES achieves significant gains over recent methods such as ViA [45], surpassing them by 5.1% and 3.7% on SmartHome (Cross-Subject) and SmartHome (Cross-View2), respectively. For lab-setting actions, we transfer representations from NTU-RGB+D 120 to the smaller PKU-MMD II dataset and report results in Table IV, where SKEYES outperforms the state-of-the-art (SoTA), highlighting its strong transferability across both lab-controlled and real-world environments. These results demonstrate the effectiveness of our dual-perspective design in capturing complementary and mutually reinforcing information, resulting in more discriminative representations across action classes.

c) Action Classification (Semi-supervised Evaluation):

Evaluating model generalizability by fine-tuning a pretrained model with limited labeled data is essential, especially for real-world datasets where annotations are costly and scarce. In Table VI and V, we report results using only 5% and 10% of labeled samples from both the lab-setting dataset NTU-RGB+D 60 and the real-world datasets SmartHome and Penn Action. SKEYES generalizes well across different datasets and achieves SoTA performance, which demonstrates the effectiveness of our design even with limited data.

Methods	Posetics		Penn Action	
	Top-1(%)	Top-1(%)	Top-1(%)	Top-1(%)
HiCo [11]	9.6	-	70.8	-
PCM ³ [50]	7.6	-	73.7	-
ViA [45]	-	-	76.9	-
USDRL [38]	12.7	-	76.7	-
SKEYES (Ours)	13.6	13.6	78.4	78.4

TABLE VII

ACTION RETRIEVAL (KNN EVALUATION) ON POSETICS AND PENN ACTION.

Method	NTU-RGB+D 60		NTU-RGB+D 120	
	CS (%)	CV (%)	CS (%)	Cset (%)
HiCLR [49]	60.6	75.1	-	-
SkeAttnCLR [14]	69.4	67.8	46.7	58.0
HiCo [11]	68.3	84.8	56.6	59.1
MAMP [26]	62.0	70.0	51.8	56.1
A ² MC [40]	70.8	85.4	59.1	62.6
PCM ³ ++ [51]	75.4	-	64.5	-
ActCLR+ [24]	75.9	81.6	58.9	62.3
SKEYES (Ours)	76.6	86.9	63.6	65.7

TABLE VIII

ACTION RETRIEVAL (KNN EVALUATION) ON NTU-RGB+D 60 & NTU-RGB+D 120.

d) Action Retrieval (KNN Evaluation): Table VII and Table VIII highlights the action retrieval performance (via KNN evaluation) on real-world and lab setting datasets, respectively. SKEYES outperforms SoTA across all evaluation protocols and shows consistent gains. Previous methods [24], [40] perform well in individual settings, but they fall short in cross-dataset generalization. In contrast, SKEYES maintains high accuracy across all protocols, indicating strong and transferable feature representations. These results further validate SKEYES’s capability in learning discriminative, generalizable representations for skeleton-based action retrieval through intra-class discrimination and inter-class alignment strategies.

e) Action prediction (Linear Evaluation): In Table X, we report action prediction results on real-world datasets, including SmartHome-CV2, Penn Action, and Posetics, as well as the lab-controlled NTU-RGB+D60 (CS) benchmark, following the protocol of [35], assessing the model’s ability to recognize an action as early as possible by observing only a portion of the skeleton sequence (70%, 80%, or 90%) under the same setting of linear evaluation. This task involves masking part of the skeleton sequence and evaluating on the observed part to recognize the action. SKEYES consistently outperforms prior methods by a clear margin, highlighting its robustness to real-world noise, viewpoint variations, and diverse action distributions.

f) Action Detection (Transfer Learning): As activities in real-world are typically collected in untrimmed and temporally continuous videos, it is essential to evaluate the model in the context of action detection. In Table IX, we report action detection results on PKU-MMD I dataset under a linear evaluation protocol, using mean Average Precision (mAP) at different temporal Intersection over Union (tIoU) thresholds: 0.1, 0.3, and 0.5. For fair comparison with (SoTA) methods, we use an encoder pretrained on NTU-RGB+D 60 (cross-subject) and attach a linear classifier to predict frame-level action categories, thereby generating the final proposals. The results demonstrate the effectiveness

Method	mAP@tIoU (%)		
	0.1	0.3	0.5
AimCLR [33]	43.9	-	35.1
HiCo-Transformer [11]	32.5	31.8	28.6
HiCo-GRU [11]	50.1	48.6	44.3
SkeAttnCLR [14]	48.5	-	41.7
LAC [47]	55.2	-	-
SCDNet [39]	65.6	64.7	58.5
SKEYES (Ours)	73.6	72.4	65.0

TABLE IX
ACTION DETECTION (TRANSFER LEARNING "LINEAR") ON
PKU-MMD I.

Method	70%	80%	90%	100%
<i>SmartHome-CV2</i>				
HiCo [11]	38.2	42.9	45.3	54.8
PCM ³ [50]	21.2	24.4	29.3	46.8
USDRL [38]	48.2	51.2	51.5	53.9
SKEYES (Ours)	49.0	52.4	54.1	56.9
<i>Penn Action</i>				
HiCo [11]	77.8	80.6	83.4	87.6
PCM ³ [50]	53.4	56.7	62.4	85.7
USDRL [38]	85.7	86.9	87.9	89.6
SKEYES (Ours)	86.1	88.3	90.0	91.0
<i>Posetics</i>				
HiCo [11]	18.5	19.9	20.5	21.3
PCM ³ [50]	17.1	18.1	18.7	20.0
USDRL [38]	24.2	24.8	25.4	25.9
SKEYES (Ours)	25.3	26.5	27.0	27.2
<i>NTU-RGB+D60 CS</i>				
DeepSCN [17]	58.2	60.2	60.0	58.6
MSRNN [13]	63.9	67.4	68.9	69.2
P-TSL [36]	77.6	80.1	81.5	82.0
SCDNet [39]	79.5	82.1	85.0	86.6
USDRL [38]	80.3	82.4	83.7	84.2
SKEYES (Ours)	81.6	84.4	86.1	87.2

TABLE X
ACTION PREDICTION- *Linear Evaluation results-* ON
REAL-WORLD DATASETS (SMARTHOME-CV2, PENN ACTION
AND POSETICS) TOP AND ON LAB-SETTING DATASETS
(NTU-RGB+D60 CS) BOTTOM.

of our dual-framework approach, which not only enhances performance but also outperforms traditional self-supervised learning (SSL) methods [47], [11], [39].

V. ABLATION STUDY

a) Which Models are Complementary?: Since our framework is modular and plug-and-play, it allows for flexible integration of various backbone models. We conducted an ablation study on different GCN and transformer backbones, including topology-based methods [5], [7], [18] where uses a fixed or guided graph structure, and Topology-free [44], [53], [10] that learn joint relationships implicitly, without relying on any prior skeletal connectivity on NTU-RGB+D 60 dataset (see Table XI). Each backbone encodes the skeleton sequence differently to capture useful information. This study aims to identify which models produce complementary features that, when combined, enhance the understanding of human activities from skeleton representations. As shown in Table XI, homogenized GCN models such as CTR-CTR,

Backbone	CS (%)	Pos (%)	Backbone	CS (%)	Pos (%)
CTR - CTR	87.2	26.5	CTR - UNIK	86.6	27.2
CTR - Info	85.9	25.0	CTR - HD	85.1	24.1
CTR - Hyper	86.1	25.5	UNIK - Info	83.9	21.3
UNIK - HD	84.8	23.2	UNIK - Hyper	83.4	21.2
Info - Hyper	83.5	22.6	HD - Hyper	85.2	22.4

TABLE XI
ABLATION STUDY USING DIFFERENT BACKBONES ON
NTU-RGB+D 60-CS AND POSETICS (POS)

Index	DACo		NTU-RGB+D 60	
	(Contrastive)	(Alignment)	CS(%)	CV(%)
A0:	One Perspective [39]	×	86.6	91.7
A1:	×	MSE	60.5	65.3
A2:	Mix(1&2)	×	80.1	85.6
A3:	Both w/ Mix(1&2)	×	84.2	91.6
A4:	Both w/ Mix1	MSE	86.1	91.5
A5:	Both w/ Mix2	MSE	86.7	89.3
A6:	Both w/ Mix(1&2)	MSE	85.1	92.1
A7:	Both w/o Mix(1&2)	MSE	87.2	92.3

TABLE XII
IMPACT OF DUAL ALIGNMENT ON CONTRASTIVE LOSS ON
NTU-RGB+D 60.

as well as un-homogenized models like CTR-UNIK, are particularly effective when used together. This is because they capture diverse and complementary information from the same skeleton sequences. As previous SOTA methods [39], [6] have shown that topology-based methods are the most effective on laboratory datasets, we use CTR-CTR as backbones in these settings. Differently for real-world datasets which are more complex and present many challenges like occlusions, the use of CTR-UNIK to leverage both topology-based and topology-free backbones is the most effective for these scenarios [45].

b) Impact of Dual Alignment on Contrastive Loss:

With our dual-branch design, we get two different feature views from the same skeleton sequence. As shown in Figure 3, we can either align these features directly by minimizing their distance or apply a contrastive loss that mixes positives and negatives across the two branches. Table XII (Grey rows) shows that using only feature alignment doesn't perform well, as minimizing distance alone lacks the ability to learn discriminative features [A1]. On the other hand, using contrastive loss alone also falls short due to a mismatch between the branches, which causes issues when mixing samples [A2, A3]. Combining both alignment and contrastive learning helps close this gap. It improves how positive and negative pairs are formed and makes the model learn more

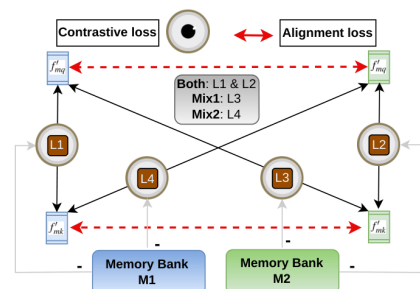


Fig. 3. Illustration of Dual Alignment Contrastive space.

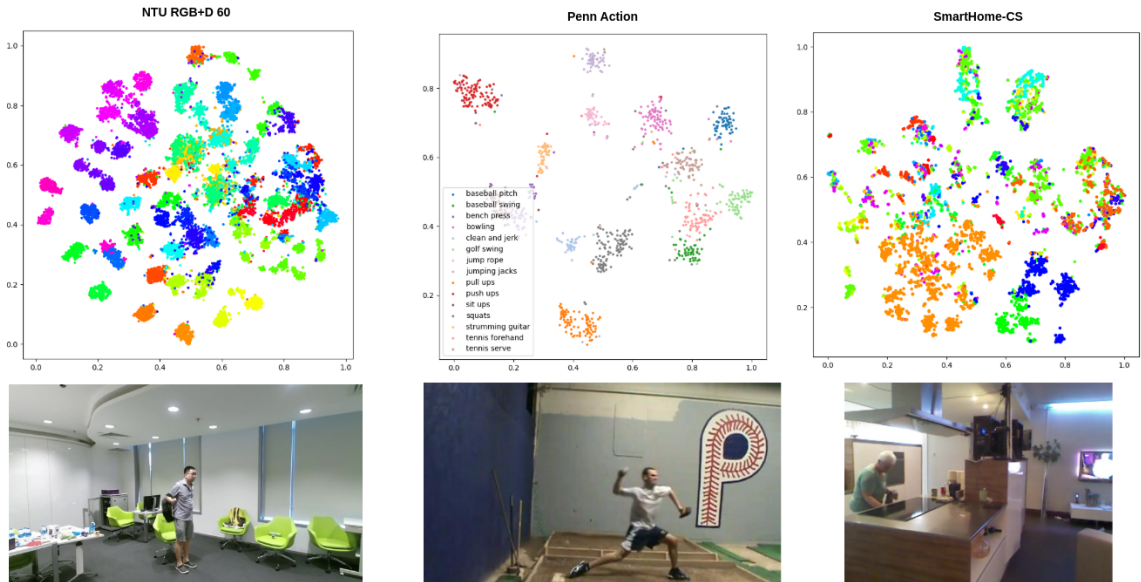


Fig. 4. TSNE plots of SKEYES’s representation on NTU-RGB+D 60 CV, Penn Action and SmartHome-CS datasets.

SmartHome-CV2	
Activity	SKEYES Gain
Usetablet	+64.67
Pour.Frombottle	+28.77
Eat.Snack	+19.43
Cutbread	+17.05
Walk	+15.20
Getup	-1.27
Usetaptop	-5.85
Pour.Fromcan	-6.66
Mean	+3.7

SmartHome-CS	
Activity	SKEYES Gain
Drink.Fromcup	+44.15
Leave	+42.56
Readbook	+34.01
Usetaptop	+28.60
Pour.Fromkettle	+29.85
Cook.Cut	-5.24
Pour.Fromcan	-6.64
Maketea.Boilwater	-8.55
Mean	+5.1

TABLE XIII

GAIN FROM SKEYES ACROSS DIFFERENT ACTIVITIES FOR SMARTHOME DATASET COMPARING WITH ViA [45]

meaningful representations. As seen in Table XII [A5, A6], this combination leads to strong results with or without feature mixing. The best and most generalizable setting applies contrastive loss within each branch and adds alignment between them [A7]. This setup achieves consistently high performance with less training overhead, since mixing losses are not required.

c) Deeper Insights, Analysis and Comparisons: In Table XIII, we provide a per-class performance analysis in comparison to state-of-the-art methods across SmartHome dataset. More analysis is reported in the Appendix. The results demonstrate the effectiveness of our framework in improving class-wise accuracy. SKEYES yields significant gains on actions such as “Use tablet”, which typically involve long-duration or fine-grained motion cues. These actions benefit from our dual-perspective design: two complementary spatio-temporal encoders (e.g., topology-guided CTR and topology-free UNIK) capture diverse motion cues, whose alignment leads to more discriminative representations than view-invariant methods such as ViA [45]. The model shows consistently high performance on the NTU-RGB+D 60 and Penn Action datasets. However, in the SmartHome dataset, decreases in actions like “Pour from can” suggest that skeleton-only representations remain insufficient when object-

level appearance is the key discriminative cue. We will work to extend SKEYES in future to multi-modal settings.

d) Representation Visualization Using t-SNE: We use t-SNE plots to illustrate SKEYES’s representation learning capability across various scenarios, including laboratory, activity daily living (ADL), and sports settings. As shown in Figure 4, the learned representations are well-separated in NTU-RGB+D 60 and Penn Action, which feature fewer challenges. In contrast, the ADL SmartHome dataset shows more overlap due to factors e.g., occlusions, cluttered environments, and the presence of both short and long actions.

VI. CONCLUSION

We present SKEYES, a novel dual-perspective self-supervised framework for skeleton-based action representation learning. By decoupling spatial and temporal modeling through dual encoders and introducing a dual contrastive learning objective, SKEYES effectively captures fine-grained motion patterns and robust representations. Extensive evaluations on a large number of benchmark datasets demonstrate that SKEYES achieves state-of-the-art performance across action classification, detection, and transfer learning tasks, with strong generalization under real-world and low-label conditions.

REFERENCES

- [1] C. Caetano, F. Br mond, and W. Schwartz. Skeleton image representation for 3D action recognition based on tree structure and reference joints. *SIBGRAPI*, 2019.
- [2] W. Cao, L. Qian, Y. Zhang, X. Li, and X. Yin. Asymmetric context-guided adaptive alignment network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- [3] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning, 2020.
- [4] 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 *ICCV*, 2021.
- [5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021.
- [6] Y. Chen, L. Zhao, J. Yuan, Y. Tian, Z. Xia, S. Geng, L. Han, and D. N. Metaxas. Hierarchically self-supervised transformer for human skeleton representation learning. In *ECCV*, 2022.
- [7] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [8] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, and L. Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv:1703.07475*, 2017.
- [9] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca. Toyota smarhome: Real-world activities of daily living. In *ICCV*, 2019.
- [10] J. Do and M. Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, 2024.
- [11] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *AAAI*, 2023.
- [12] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. In *CVPR*, 2022.
- [13] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang. Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [14] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu. Part aware contrastive learning for self-supervised action recognition, 2023.
- [15] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, and D. Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 34(11):9389–9403, 2022.
- [16] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.
- [17] Y. Kong, Z. Tao, and Y. Fu. Deep sequential context networks for action prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [18] J. Lee, M. Lee, D. Lee, and S. Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10444–10453, October 2023.
- [19] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang. 3d human action representation learning via cross-view consistency pursuit. In *CVPR*, 2021.
- [20] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang. 3d human action representation learning via cross-view consistency pursuit. In *(CVPR)*, 2021.
- [21] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021.
- [22] L. Lin, J. Zhang, and J. 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.
- [23] L. Lin, J. Zhang, and J. Liu. Mutual information driven equivariant contrastive learning for 3d action representation learning. *IEEE Transactions on Image Processing*, 2024.
- [24] L. Lin, J. Zhang, and J. Liu. Self-supervised skeleton representation learning via actionlet contrast and reconstruct. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [25] 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.
- [26] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li. Masked motion predictors are strong 3d action representation learners. In *ICCV*, 2023.
- [27] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *ECCV*, 2022.
- [28] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3D human activity analysis. *CVPR*, 2016.
- [29] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with directed graph neural networks. *CVPR*, 2019.
- [30] J. J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam, and T. Liu. View-invariant probabilistic embedding for human pose. In *(ECCV)*, 2020.
- [31] S. Sun, D. Liu, J. Dong, X. Qu, J. Gao, X. Yang, X. Wang, and M. 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.
- [32] F. M. Thoker, H. Doughty, and C. G. M. Snoek. Skeleton-contrastive 3d action representation learning, 2021.
- [33] G. Tianyu, L. Hong, C. Zhan, L. Mengyuan, W. Tao, and D. Runwei. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI*, 2022.
- [34] H. Wang, X. Ma, J. Kuang, and J. Gui. Heterogeneous skeleton-based action representation learning. In *CVPR*, 2025.
- [35] H. Wang, W. Weng, J. Wang, F. Zhao, G. sen Xie, X. Geng, and L. Wang. Foundation model for skeleton-based human action understanding. *Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [36] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng. Progressive teacher-student learning for early action prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [37] W. Wei, S. Zhang, Y. Dang, and J. Yin. Masksem: Semantic-guided masking for learning 3d hybrid high-order motion representation. In *IROS*, 2025.
- [38] W. Weng, H. Wang, J. Wang, L. He, and G. Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [39] C. Wu, X.-J. Wu, J. Kittler, T. Xu, S. Atito, M. Awais, and Z. Feng. Sd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *AAAI*, 2024.
- [40] B. Xu, X. Shu, J. Zhang, R. Yan, and G.-S. Xie. Attack-augmentation mixing-contrastive skeletal representation learning, 2024.
- [41] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.
- [42] D. Yang, R. Dai, Y. Wang, R. Mallick, L. Minciullo, G. Francesca, and F. Bremond. Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos. In *WACV*, 2021.
- [43] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond. Self-supervised video pose representation learning for occlusion-robust action recognition. In *FG*, 2021.
- [44] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond. Unik: A unified framework for real-world skeleton-based action recognition. In *BMVC*, 2021.
- [45] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond. Via: View-invariant skeleton action representation learning via motion retargeting. *IJCV*, 2024.
- [46] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Br mond. Self-supervised video pose representation learning for occlusion-robust action recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021.
- [47] D. Yang, Y. Wang, A. Dantcheva, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond. Lac - latent action composition for skeleton-based action segmentation. In *ICCV*, 2023.
- [48] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *ICCV*, 2021.
- [49] J. Zhang, L. Lin, and J. Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [50] J. Zhang, L. Lin, and J. Liu. Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In *Proceedings of the ACM International Conference on Multimedia*, 2023.
- [51] J. Zhang, L. Lin, S. Yang, and J. Liu. Self-supervised skeleton-based action representation learning: A benchmark and beyond. *IJCV*, 2026.
- [52] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding.

- In *ICCV*, 2013.
- [53] Y. Zhou, Z.-Q. Cheng, C. Li, Y. Geng, X. Xie, and M. Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.
 - [54] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023.
 - [55] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *CVPR*, 2020.