# Representation-Centric Survey of Skeletal Action Recognition and the ANUBIS Benchmark

Yang Liu, Jiyao Yang, Madhawa Perera, Pan Ji, Dongwoo Kim, Min Xu, Tianyang Wang, Saeed Anwar, Tom Gedeon, Lei Wang, Zhenyue Qin

***Abstract*—3D skeleton-based human action recognition has emerged as a powerful alternative to traditional RGB and depth-based approaches, offering robustness to environmental variations, computational efficiency, and enhanced privacy. Despite remarkable progress, current research remains fragmented across diverse input representations and lacks evaluation under scenarios that reflect modern real-world challenges. This paper presents a representation-centric survey of skeleton-based action recognition, systematically categorizing state-of-the-art methods by their input feature types: joint coordinates, bone vectors, motion flows, and extended representations, and analyzing how these choices influence spatial-temporal modeling strategies. Building on the insights from this review, we introduce ANUBIS, a large-scale, challenging skeleton action dataset designed to address critical gaps in existing benchmarks. ANUBIS incorporates multi-view recordings with back-view perspectives, complex multi-person interactions, fine-grained and violent actions, and contemporary social behaviors. We benchmark a diverse set of state-of-the-art models on ANUBIS and conduct an in-depth analysis of how different feature types affect recognition performance across 102 action categories. Our results show strong action-feature dependencies, highlight the limitations of naïve multi-representational fusion, and point toward the need for task-aware, semantically aligned integration strategies. This work offers both a comprehensive foundation and a practical benchmarking resource, aiming to guide the next generation of robust, generalizable skeleton-based action recognition systems for complex real-world scenarios. The dataset website, benchmarking framework, and download link are available at https://yliu1082.github.io/ANUBIS/.**

***Index Terms*—Skeleton-based action recognition, Human activity understanding, Feature fusion, Representation learning, Joint-bone-motion features, Benchmark, ANUBIS, Multi-person interaction, Privacy-preserving vision, Spatio-temporal modeling.**

## I. Introduction

HUMAN action recognition from visual data represents a fundamental challenge in computer vision, requiring robust extraction and analysis of spatio-temporal patterns from complex visual sequences [5], [6], [15], [16], [55], [91], [93], [99], [104], [106]. The ability to automatically classify human activities from video data has enabled critical applications across intelligent surveillance systems [63], [94], autonomous vehicle perception [36], robotic interaction [1], augmented reality interfaces [4], clinical motion analysis [92], and behavioral monitoring [59], [131]. To achieve robust action recognition, researchers have explored diverse data modalities [6], [16], [91], [93]. RGB videos provide rich visual cues such as appearance, texture, and color, effectively capturing the overall dynamics of human movement and contextual information.

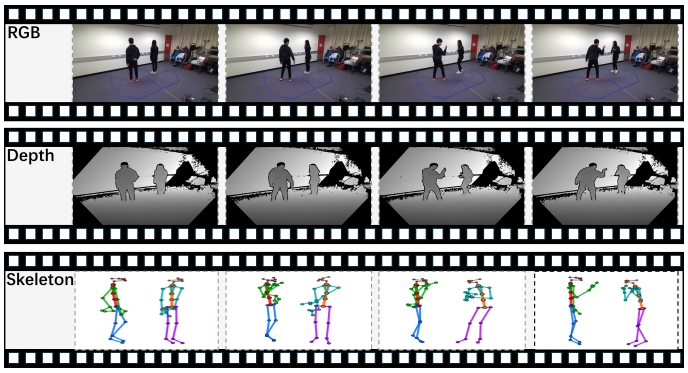



Fig. 1: Multi-modality example of the action *wave knife to others* from the ANUBIS dataset, captured uniquely from a back-view perspective. Four consecutive frames are shown across RGB (top), Depth (middle), and 3D Skeleton (bottom) modalities. This example illustrates ANUBIS's distinctive contributions in introducing previously unseen interaction classes and incorporating back-view acquisition, both absent in prior skeleton-based action datasets (see Sec. III).

However, RGB-based methods face significant limitations: they suffer from performance degradation under challenging environmental conditions, including poor lighting, background clutter, and appearance variations, and as dense data, RGB videos are computationally intensive, typically requiring larger models and substantial computational resources for processing [7], [92], [100], [109], [120]. These limitations collectively restrict their reliability and feasibility in real-world deployments. Depth maps offer complementary 3D structural information that enables better geometric and spatial modeling, yet depth-based methods remain sensitive to viewpoint changes, occlusions, and sensor noise [26], [91], [93], [108]. Infrared data, while resilient to lighting variations, presents challenges for infrared-based methods, which often lack semantic richness and struggle to capture fine-grained motion details [83]. Given the limitations of these modalities, 3D human skeleton data, which represents human poses using a sparse set of key anatomical joints, has

Yang Liu and Jiyao Yang contributed equally to this work.

Y. Liu is with the Australian National University. J. Yang is with the University of Alabama at Birmingham. M. Perera was with the Australian National University, now is with Data61 CSIRO. P. Ji was with OPPO US Research Center. D. Kim is with POSTECH. M. Xu is with Carnegie Mellon University. T. Wang is with the University of Alabama at Birmingham. S. Anwar was with both Australian National University (ANU) and Data61, CSIRO, now is with the University of Western Australia. T. Gedeon was with the Australian National University, now is with Curtin University.

L. Wang is with the School of Engineering and Built Environment, Electrical and Electronic Engineering, Griffith University, and with Data61/CSIRO (e-mail: l.wang4@griffith.edu.au). Z. Qin is with the School of Computing, Australian National University, and School of Medicine, Yale University (e-mail: zhenyue.qin@yale.edu).

Lei Wang and Zhenyue Qin are corresponding authors.

emerged as a highly effective representation for human action analysis [34], [66], [96]–[98], [102], [103].

Compared to other modalities, skeleton data offer several advantages (see Fig. 1): significantly reduced storage requirements due to their sparse nature, computational efficiency during processing, robustness to environmental variations, invariance to appearance changes, and enhanced privacy protection by removing personally identifiable visual features [20], [21], [56], [92]. These qualities make skeleton data particularly well-suited for deployment in challenging real-world scenarios, edge devices with limited computational resources, and privacy-sensitive applications [11], [59], [68], [105]. With recent advances in depth sensing technologies and pose estimation algorithms (e.g., RGB-D cameras, LiDAR sensors, and deep learning-based pose estimators), acquiring high-quality skeleton sequences has become increasingly accessible. These advantages have driven significant advances in skeleton-based action recognition across three key dimensions. Input representation has evolved beyond joint coordinates to incorporate bone vectors, velocities, accelerations, and surface normals for richer motion encoding. *Spatial modeling* has been transformed by GCNs, with architectures like ST-GCN [118], 2s-AGCN [76], and DeGCN [58] effectively capturing skeletal topology. *Temporal modeling* has progressed from recurrent networks to advanced spatio-temporal convolutions (e.g., MS-G3D [54]) and Transformers for long-range dependencies [103]. Attention mechanisms [6], cross-modal fusion [13], and neural architecture search have further enhanced the performance [100].

Equally critical to this progress has been the availability of large-scale datasets, particularly the NTU RGB+D series [51], [73], which have served as foundational benchmarks for the field. However, these datasets increasingly fall short of meeting evolving research demands in several key aspects. First, the majority of actions are captured from frontal viewpoints with limited pose diversity, particularly *lacking back-view perspectives from multiple angles*, which constrains model robustness when deployment scenarios involve varied camera angles or when subjects are oriented away from the primary sensor. Second, existing datasets *predominantly focus on individual daily activities while neglecting complex multi-person interactions*, such as handshaking and collaborative behaviors, which are prevalent in real-world environments. Third, they *fail to incorporate challenging actions involving aggression and violence* (e.g., hitting someone's head, stabbing with weapons), which represent important categories for security and surveillance applications but remain largely absent from current benchmarks. Finally, they *lack contemporary socially relevant behaviors* (e.g., pandemic-related gestures and social distancing protocols), thereby limiting their ecological validity and applicability to modern scenarios. These gaps collectively highlight the urgent need for more comprehensive datasets that better reflect the complexity and diversity of human actions in real-world.

In this work, we address key limitations in current skeleton-based action recognition through three main **contributions**:

i. **Representation-centric survey.** We provide a systematic taxonomy of skeleton-based action recognition methods, organized by input representation (joints, bones, motion, extended features), and analyze how spatial-temporal modeling strategies adapt to each representation type.

ii. **ANUBIS benchmark dataset.** We propose ANUBIS, a large-scale benchmark of 102 diverse actions. ANUBIS uniquely incorporates: (i) multiple viewpoints including back-view recordings, (ii) complex multi-person interactions, (iii) challenging violent and security-critical actions,filling critical gaps in existing datasets.

iii. **Comprehensive benchmarking and analysis.** We evaluate a wide range of popular models on ANUBIS, showing how representation choice and modeling strategy affect performance, and uncovering cases where naïve multi-representational fusion degrades recognition.

These contributions collectively position ANUBIS as a challenging new benchmark, deepen understanding of representation-driven design, and provide a foundation for developing more generalizable and semantically aware skeleton-based action recognition systems.

## II. A Representation-Centric Review

### A. Joint-Based Methods

Joint coordinates are characterized by multi-node structures, where the node count is typically determined by the skeletal detection sensors used [91], [93]. Each node encodes the spatial position of a human joint through either two-dimensional pixel coordinates or three-dimensional world coordinates, forming the geometric primitives of the skeletal graph structure.

*1) Spatial Modeling:* Effective human pose analysis requires capturing both individual joint positions and their inherent spatial relationships. While joint coordinates provide explicit 2D/3D locations, the underlying skeletal structure is defined by connectivity and dependency patterns among joints. Spatial modeling approaches for joint coordinates can be categorized into two primary strategies: methods using predefined anatomical connectivity and adaptive approaches that learn joint relationships from data.

**Manually predefined joint connectivity.** Predefined connectivity methods embed fixed topological relationships among joints directly into model architectures using binary adjacency matrices that encode anatomical connections. Unlike preprocessed bone methods that introduce 3D geometric vectors as additional input features, predefined connectivity serves as structural constraints within the architecture itself. This architectural approach offers computational efficiency since connectivity patterns remain static throughout training and inference, requiring negligible overhead. Early CNNs establish structured feature representations by using fixed skeletal connections within convolutional operations, where local receptive fields extract spatially-constrained features according to anatomical topology [18]. Contemporaneously, RNNs perform spatial modeling through sequence conversion using two primary strategies: sequential linearization transforms joint coordinates into structured orderings (chain, traversal, or tree sequences) that preserve local connectivity and spatial adjacency, while hierarchical partitioning uses two-layer architectures to capture both part-level details and whole-level structural integrity [19], [52], [53], [74], [87], [91], [93], [125]. Advancing beyond these sequential approaches, GNNs use predefined adjacency matrices to constrain graph convolutions, enabling feature aggregation

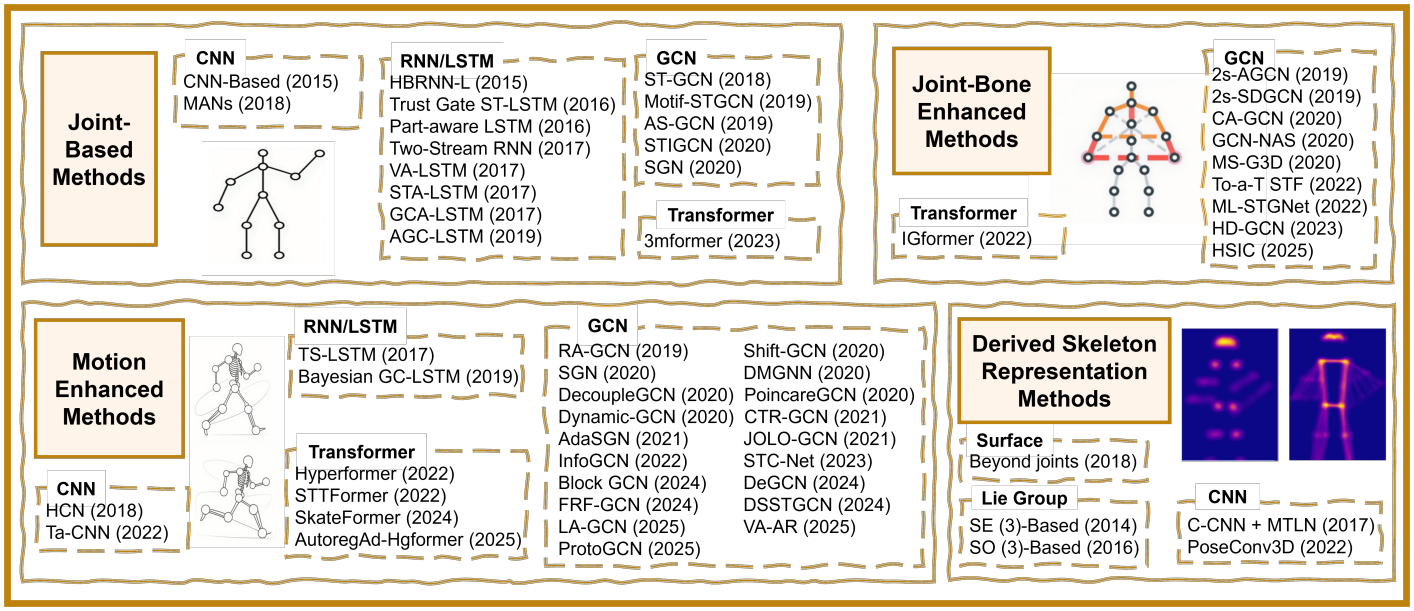



Fig. 2: Evolution of skeleton-based action recognition methods from 2014 to 2025. The taxonomy categorizes approaches into four main groups: Joint Based (joint only), Join-Bone Enhanced (ioint+bone), Motion Enhanced (joint+bone+motion), and Derived Skeleton Representation Methods. Each category showcases the progression from traditional RNN/LSTM methods to modern GCN and Transformer architectures, demonstrating the evolution of deep learning techniques for skeleton-based action recognition.

only among directly connected joints while constructing tree-like hierarchies based on parent-child relationships. These methods often convert absolute coordinates to local representations relative to parent nodes, eliminating global translation effects and incorporating geometric constraints to ensure physiological plausibility [118], [128].

However, traditional graph structures with binary edge connections prove insufficient for modeling complex multi-joint coordination behaviors [92], [98]. Consequently, hypergraph extensions address this limitation through hierarchical hyperedges, first-level edges maintain original skeletal connections while higher-level edges aggregate multiple limb joints, enabling multi-branch parallel processing that extracts local motion features from low-order branches and coordination patterns from high-order branches through unified matrix token conversion [24], [71], [98], [107], [129]. While these predefined approaches collectively reduce computational overhead by avoiding preprocessed bone data, they require dataset-specific redefinition for cross-domain applications, limiting their generalizability [93], [100], [103].

**Dynamic adaptive joint relationship learning.** While manually predefined skeletal connections offer computational simplicity and efficiency, they are inherently limited to modeling relationships between adjacent joints, thereby constraining their capability for capturing complex non-adjacent dependencies essential for sophisticated action recognition. To overcome these limitations, adaptive learning approaches have emerged where models automatically learn optimal connection patterns during training, using this learned information to enhance spatial feature representation beyond fixed topological constraints.

Adaptive relationship learning provides two primary advantages: (i) Enhanced spatial modeling scope enables learning of both local adjacent relationships and global non-adjacent dependencies between arbitrary joint pairs, transcending physical connectivity limitations. For instance, in clapping gestures, bilateral hand coordination requires modeling cross-body relationships that extend far beyond skeletal connections. (ii) Action-specific feature emphasis enables models to dynamically weight joint relationships based on their discriminative importance for specific actions. For example, in running motions, lower limb joints carry substantially greater semantic significance than upper limb joints, requiring adaptive emphasis on leg-related connections.

Dynamic adaptive joint relationship learning can be achieved through three principal paradigms: attention-based adaptive weighting [98], specialized convolutional architectures [34], [96], [97], and learnable graph topology modification [13].

*Attention-based adaptive weighting.* Attention mechanisms enable models to dynamically compute relationship strengths between all joint pairs, effectively learning a fully-adaptive adjacency structure [53], [79], [80], [114]. These mechanisms have been extensively deployed in recurrent architectures including RNNs and LSTMs. Spatial attention computes functional association strengths between joint pairs based on their feature representations during forward propagation, generating data-driven connection weights that can span arbitrary spatial distances. This mechanism endows models with fully-adaptive topological modeling capabilities, autonomously discovering and reinforcing critical joint interactions across different motions. Transformer-based architectures exemplify this paradigm through sophisticated self-attention implementations [6], [17], [24], [62], [67], [71], [98], [105], [129]. These models treat skeletal structures as fully-connected graphs where every joint can potentially influence every other joint. They use Query-Key-Value mechanisms to calculate pairwise correlation scores, dynamically determining which joint relationships are most relevant for the current input. While physical skeletal connectivity can be incorporated as positional bias or structural prior, the

learned attention weights are free to deviate from anatomical constraints, enabling discovery of semantic relationships (e.g., hand-foot coordination) that transcend physical adjacency.

*Specialized convolutional architectures.* CNN-based approaches use channel-wise convolution to achieve global joint interaction modeling [41], [117]. This paradigm represents skeletal sequences as tensors, then transposes them to a new tensor format where each joint becomes a distinct input channel. Subsequent $1 \times 1$ convolution across the channels implement learnable linear combinations of all joint features, effectively enabling each spatial location to aggregate information from all joints simultaneously. This channel-wise aggregation mechanism inherently captures global dependencies since each output feature incorporates weighted contributions from all input joints, regardless of their physical adjacency.

*Learnable graph topology modification.* GCN-based approaches use learnable parameters to modify connections within graph structures. These methods typically operate on predefined adjacency matrices that encode human skeletal topology, but augment them with trainable components to enable adaptive relationship learning. Two primary strategies exist within this paradigm: (i) Adjacency matrix augmentation methods [27], [111] learn additive or multiplicative modifications to fixed adjacency matrices, enabling models to strengthen existing connections or establish entirely new pathways between previously unconnected joints. (ii) Learnable mask matrices share identical dimensionality with base adjacency matrices, where each trainable element modulates the corresponding connection strength [126]. During training, these parameters adapt to emphasize discriminative joint relationships while potentially discovering cross-limb or non-adjacent dependencies that complement the anatomical structure. Advanced variants [13], [42] can learn sparse attention patterns that selectively activate non-adjacent connections based on action-specific requirements, effectively expanding the receptive field beyond immediate neighborhoods while preserving beneficial structural priors.

*2) Temporal Modeling:* Skeleton-based action recognition fundamentally relies on capturing continuous joint movement patterns over time. Temporal modeling transforms static joint coordinates into dynamic motion trajectories by analyzing action boundaries, velocity variations, and sequential dependencies. This temporal analysis serves two critical functions: aggregating local motions across frames to form complete action semantics, thereby resolving single-frame ambiguities; and enabling discrimination between spatially similar actions (e.g., putting on versus taking off clothing) through temporal pattern analysis [102], [103], [105]. Temporal modeling faces inherent structural differences compared to spatial approaches [65], [99], [104].

Spatial modeling uses stable joint topology within individual frames, where joint relationships form consistent patterns that can be reliably captured through single-frame analysis. In contrast, temporal modeling depends on inter-frame dynamics spanning multiple key frames, where action semantics emerge from continuous motion sequences rather than instantaneous spatial configurations. This temporal dependency introduces three challenges. First, sampling-related issues arise from the difficulty of capturing complete action sequences [14]: (i) Sparse frame sampling leads to incomplete action coverage, varying action rhythms cause semantic dilution, and short-duration movements are difficult to capture adequately. (ii) Fixed sampling strategies further compound these problems, producing redundant frames for rapid motions and insufficient coverage for slower actions across different individuals and action types. Second, robustness challenges emerge from noise and occlusion disturbances that propagate through sequential processing, creating information discontinuity, instability, and temporal imbalance that significantly constrain recognition performance [105]. Third, computational challenges include long-range dependency attenuation, where local convolution operations dilute correlations between distant frames, and increased computational complexity from large receptive field requirements for complete action coverage.

Current approaches address these challenges through decomposed short-term and long-term modeling strategies for micro-actions and periodic sequences, respectively [37]. Effective solutions require preserving cross-frame joint correlations for spatio-temporal coherence, implementing adaptive time scales for motion variation accommodation, and emphasizing key frames for stable state learning. However, despite these advances, fundamental limitations persist in long-range dependency modeling and computational efficiency, necessitating continued methodological development in temporal modeling approaches. Based on the network structure, we categorize skeleton-based temporal methods into RNNs and spatio-temporal CNNs.

**RNNs.** RNNs process skeleton sequences through sequential frame input, fusing current observations with memory states encoding historical information. Standard RNNs [114] [19] exhibit gradient vanishing/explosion problems that limit long-range temporal modeling capability, making LSTMs [52] [74] [87] [125] [80] [53] [78] the preferred choice for skeleton-based recognition systems. LSTM enhancements target two primary objectives: hierarchical modeling and global context integration. Hierarchical approaches partition human joints into body parts, with part-specific LSTMs capturing local dynamics before higher-level LSTMs model inter-part temporal relationships [37] [74] [87]. Global context methods introduce temporal attention mechanisms that dynamically weight time steps and integrate historical features through context vectors, enhancing long-range dependency capture [80] [53]. These improvements significantly enhance both dependency modeling and interpretability.

**Spatio-temporal CNNs.** Temporal convolution applies one-dimensional kernels along the time axis to aggregate adjacent frame features, using dilated convolution to expand receptive fields for long-range dependency capture [31] [82]. However, pure temporal convolution ignores spatial joint structure and requires deep networks or large dilation rates for adequate receptive fields, increasing computational costs. Consequently, research has shifted toward integrated spatio-temporal convolution that jointly models spatial correlations and temporal dynamics [111] [118] [126] [27] [42]. Subsequent improvements focus on computational efficiency and representation enhancement. Residual connections optimize gradient propagation in temporal layers, while temporal attention mechanisms guide focus toward critical action periods [114] [110] [27]. Multi-scale temporal aggregation through dilated convolution captures dependencies at varying time scales, integrating short-term and long-term segments for complex action sequence modeling.

Neural architecture search (NAS) provides automated optimization for spatio-temporal convolution design [64]. By defin-

ing search spaces encompassing temporal kernel sizes, dilation rates, and connectivity patterns, evolutionary algorithms and reinforcement learning automatically discover efficient architectures, avoiding manual design biases and achieving hierarchical temporal modeling optimization. Non-Euclidean temporal feature embedding offers alternative modeling perspectives [10], [54], [96], [97], [102], [103], [111]. Hyperbolic space mapping uses exponential distance growth properties to enhance long-range temporal dependency discrimination. This approach embeds temporal dynamics into curved spaces through logarithmic and exponential transformations, mitigating distant feature confusion in Euclidean spaces, particularly beneficial for actions requiring subtle temporal distinction.

### *B. Joint-Bone Enhanced Methods*

Joint-bone enhanced methods explicitly generate bone vectors through coordinate differences between joints during preprocessing, creating independent input features that encode limb length, direction, and anatomical connectivity [75], [76].

Spatial modeling with bone explicitly incorporates anatomical structural information as additional input features alongside joint coordinates. This approach computes bone vectors during preprocessing and feeds them as independent data streams that provide explicit relationships between joints. Bone vectors are typically calculated as directional vectors between anatomically connected joints [68], [105]. Existing joint-bone enhanced methods can be categorized into two paradigms: dual-stream architecture and information fusion architecture.

**Dual-stream architecture** uses independent two-stream networks to process joint and bone representations separately, with integration occurring at the decision level through weighted combination of prediction scores [12], [17], [32], [39], [41], [54], [62], [64], [67], [76], [112], [117], [121], [129], [130], [132]. Representative methods include 2s-AGCN [76], MSG3D [54], and Dynamic GCN [121] While this approach ensures feature-specific feature learning without cross-modal interference, it incurs significant computational overhead requiring multiple times the parameters and computation costs compared to single-stream methods, and the late fusion strategy may not capture intricate inter-modal dependencies, potentially leading to suboptimal utilization of complementary information between joint coordinates and bone structural features.

**Feature fusion architecture** integrates bone as auxiliary contextual features within a unified framework, where bone data enhances joint representations through early or intermediate fusion mechanisms [13], [62], [71], [75], [100], [127]. CA-GCN [127] incorporates bone information through context-aware mechanisms that compute attention weights to determine bone feature relevance to each joint. This context term enriches the primary joint stream by providing structural information inherent in skeleton. Unlike dual-stream approaches, this paradigm maintains a single network architecture processing the primary joint stream while incorporating bone as supplementary context, resulting in significantly lower parameter count and computational requirements. However, unified processing may lead to feature interference where different feature characteristics could result in suboptimal joint representations [95], [100], [101], and the single network may not fully exploit unique properties of each feature that dedicated streams could capture.

The spatial modeling approaches for joint-bone representations fundamentally differs from joint-only methods, including manually predefined joint connectivity and dynamic adaptive joint relationship learning, which construct bone relationships implicitly within the network. Explicit bone preprocessing offers distinct advantages over joint-only approaches: pre-computed bone vectors directly encode physical connections between adjacent and non-adjacent joints, using anatomical prior knowledge for immediate access to limb relationships and biomechanical constraints. This design concentrates computational costs in the preprocessing stage while requiring only multi-stream feature fusion during inference, contrasting with joint-only methods that must predefine or learn these relationships within the network. Each method presents characteristic trade-offs. Joint-bone enhanced methods provide computational efficiency and anatomical consistency but are limited to predefined structural relationships. Among joint-only approaches, manually predefined connectivity offers simplicity but restricts modeling to adjacent joints, while dynamic adaptive learning enables flexible capture of non-adjacent joint relationships, a key advantage for complex action modeling, but requires large-scale training data and incurs higher computational costs. The choice depends on application requirements: joint-bone enhanced methods suit scenarios requiring anatomical consistency and efficiency, while dynamic adaptive approaches excel when flexible non-adjacent joint modeling is crucial for complex actions.

### *C. Motion Enhanced Methods*

Motion represents dynamic changes extracted from temporal variations between adjacent frames, capturing kinematic dependencies and dynamic spatial information essential for distinguishing actions that remain ambiguous through spatial analysis alone. This representation encompasses two fundamental categories of kinematic information characterizing skeletal sequence temporal evolution, joint motion information and bone motion information. Joint motion information captures dynamic changes of individual joints across temporal dimensions, typically represented as coordinate differences between consecutive frames. For each joint at a given time frame, its motion feature is defined as the change in position from the previous frame, effectively capturing the joint's velocity [105]. In addition, bone motion captures the dynamic changes in the relationships between connected joints over time [66]. Specifically, it represents how the relative positions of connected joints evolve from one frame to the next. This bone motion is particularly important for modeling actions that involve complex structural changes, such as limb extensions or coordinated movements across multiple limbs, as it emphasizes the dynamics of skeletal connectivity rather than just individual joint displacement [10], [75].

Recent research adopts two primary architectures for motion integration. *Multi-stream architectures* process joint, bone, and motion features through separate network branches, subsequently integrated via concatenation or weighted fusion strategies [8], [10], [12], [38], [49], [58], [67], [75], [77], [110], [115], [116], [121], [124], [129]. Motion streams typically use more refined temporal modeling to capture dynamic characteristics compared to joint and bone streams. *Unified network approaches* directly combine motion features with joint and

bone features by concatenating along channel dimensions within single architectures [43], [71], [81], [124], [126], [127]. Both architectures benefit from two motion modeling strategies. Attention mechanisms focus on informative motion components [12], [38], [49], [58], [71], [110], [115], [116], [124], [126], with spatial attention highlighting active joint nodes, temporal attention locating keyframes within action sequences, and cross-modal attention coordinating complementary relationships between different representations. Multi-scale temporal modeling processes motion signals across different temporal resolutions [8], [12], [38], [115], [116], [124], [130], using parallel convolution branches with small kernels for fine-grained high-frequency motions and large kernels for long-period actions, preserving dynamic features at multiple temporal resolutions.

### *D. Derived Skeleton Representation Methods*

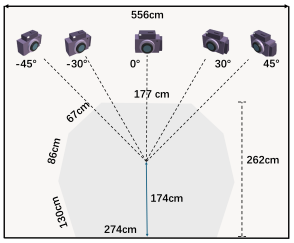
Researchers explore derived skeleton representations with enhanced expressiveness through mathematical transformations, physical model derivations, or visual processing techniques.

**Surface normal** approaches represent the geometry of body parts by capturing the relative shapes formed by adjacent edges. For each pair of connected edges, a surface normal vector describes the orientation of the corresponding body surface. To maintain consistency with the scale of joint coordinates, these vectors are appropriately scaled. For a skeleton with multiple joints, each joint defines a corresponding surface, resulting in a tensor of dimensions corresponding to the number of frames, the number of joints, and the three components of the normal vector [88]. Surface normals capture shape information during movements, including planar angle changes formed by body parts, thereby enhancing multi-joint spatial interaction representation and limb posture complexity modeling. In temporal modeling, surface normal variations effectively represent planar configuration evolution in limb movements, complementing joint and edge features in the temporal dimension. Bidirectional LSTM processing enhances time-dependent action relationship learning. However, surface normal computation faces two limitations. First, edge vector calculation errors from noise or occlusion propagate to surface normal vectors, introducing erroneous geometric dynamics in temporal modeling. Second, retaining only planes may lose potential geometric correlations between non-adjacent joints in time series, affecting complete expression of subtle temporal action features.

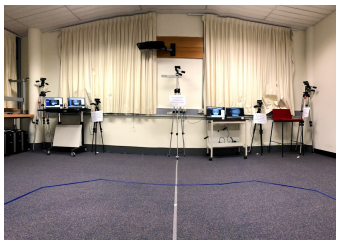
**Lie group**-based methods represent human skeletons using two main frameworks. SE(3)-based approaches model skeletons as a set of rigid body transformations, capturing both rotation and translation to comprehensively describe spatial pose relationships [85]. In contrast, SO(3)-based approaches focus on the relative rotational relationships between body parts, using only rotation matrices and applying scale normalization to retain 3D rotational information. This allows for a skeletal representation that is invariant to scale while emphasizing rotational dynamics [86]. Both SE(3)- and SO(3)-based approaches model action sequences as curves on Lie group manifolds and handle temporal variations using Dynamic Time Warping (DTW). SE(3)-based methods project action curves onto the corresponding Lie algebra, extract multi-scale temporal features through a Fourier Time Pyramid, and perform classification using a linear SVM. This framework captures dynamic patterns at different temporal resolutions, effectively addressing variations in action speed. SO(3)-based methods, on the other hand, introduce rolling mapping mechanisms that unfold the Lie group manifold along nominal action curves. This reduces distortion during mapping while preserving distance relationships, enabling more accurate modeling of rotational dynamics in skeletal motion.

Lie group representations provide geometrically meaningful skeleton modeling through structured manifold point sets. SE(3) methods preserve complete spatial pose information including both rotation and translation, while SO(3) methods achieve computational efficiency and cross-individual generalization through dimensionality reduction via rotational focus. However, these approaches face several limitations. SE(3) representations incur higher computational complexity due to increased feature dimensionality from rotation and translation components. SO(3) methods may sacrifice spatial translation information that could be relevant for certain action categories. Both approaches exhibit decreased discriminative ability when handling extreme postures or complex action couplings due to nonlinear Lie group manifold characteristics. Additionally, rolling mapping algorithms, while reducing manifold distortion, introduce implementation complexity that may limit practical deployment [91], [93].

**Joint heatmap** methods convert 2D skeleton coordinates into probabilistic spatial representations using Gaussian distributions. Each joint is represented as a heatmap that reflects both its spatial location and confidence, with the spread of the Gaussian controlling the uncertainty around the joint position. By stacking these heatmaps over time, a 3D volume is formed, capturing both spatial and temporal information of the skeleton. This volume has dimensions corresponding to the number of joints, the temporal length, and the spatial resolution, providing a rich representation for downstream tasks [22]. The probabilistic heatmap representation enables robust pose estimation under noise conditions while maintaining computational efficiency through direct multi-joint accumulation. The 3D volume structure facilitates direct 3D-CNN processing for spatiotemporal feature extraction. *Spatial modeling* uses adapted 3D-CNN architectures that remove early-stage downsampling operations to preserve low-resolution heatmap features. Shallow network designs reduce computational complexity while maintaining spatial dependency capture capabilities. Preprocessing involves minimum bounding box localization and cropping to focus on action subjects while preserving spatial relationships. Multi-channel input combining joint and limb heatmaps enhances skeletal structure representation through collaborative modeling. *Temporal modeling* uses 3D convolution kernels to simultaneously extract spatial joint relationships and temporal position changes across adjacent time steps. This unified spatiotemporal processing captures dynamic joint evolution patterns essential for action recognition. Heatmap-based representations provide natural noise robustness through probabilistic encoding and efficient 3D-CNN compatibility. The approach enables direct spatial relationship modeling without explicit graph construction while maintaining computational tractability. However, limitations include spatial resolution constraints that may lose fine-grained positional information, increased memory requirements for 3D volume storage, and potential temporal redundancy in slowly-varying actions. Additionally, the Gaussian assumption




(a) Venue layout.



(b) Camera arrangement.

Fig. 3: ANUBIS dataset collection setup overview.

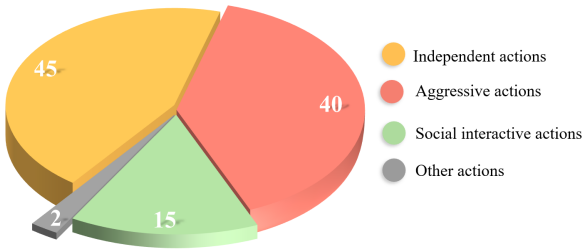



Fig. 4: Distribution of 102 human actions classified into four categories. The pie chart shows: independent actions (45, 44.1%), aggressive actions (40, 39.2%), social interactive actions (15, 14.7%), and other actions (2, 2.0%). Other actions specifically refer to spatial position change behaviors, including walk apart and walk from apart to together.

may not optimally represent all joint confidence distributions, particularly for occluded or uncertain joint detections. Below, we present our dataset along with a comparative analysis.

## III. ANUBIS: A New Benchmark Dataset

**Motivation.** The collection of ANUBIS dataset addresses three fundamental limitations in existing datasets that constrain practical deployment effectiveness. First, established datasets including NTU60 [73] and NTU120 [51] capture predominantly frontal and lateral viewpoints while systematically excluding rear-view perspectives across multiple angles, limiting model robustness when processing actions from challenging observation positions that commonly occur in real-world monitoring scenarios. Second, these benchmarks focus primarily on individual behaviors while underrepresenting multi-person interactions such as collaborative activities and assistance behaviors that are prevalent in practical applications. Third, existing datasets exhibit constrained action scope, omitting aggressive behaviors critical for security and surveillance deployment, such as stabbing and striking actions, alongside pandemic-induced social adaptations including elbow touching and arm-directed sneezing that emerged following COVID-19. These gaps make existing datasets much less useful in real-world applications, as models trained on such data cannot recognize actions from challenging camera angles, multi-person interactions, and new types of behaviors that are essential for practical systems. Thus, our new dataset overcomes these critical limitations, advancing the field toward practical real-world applicability.

### A. Dataset Collection

ANUBIS comprises 102 carefully selected actions including both individual behaviors (e.g., drinking water, waving) and multi-person interactions (e.g., handshaking, object exchange, stabbing). The 102 actions are distributed across 40 collection sessions, with each session involving two participants as a group and lasting approximately 1.5 hours. Every 10 sessions incorporates 10-minute breaks to maintain performance quality, and sessions exhibiting substandard action execution are re-recorded to ensure data integrity. The dataset comprises 40 participant groups totaling 80 participants and approximately 60 hours of multi-modal recordings (see Fig. 1).

**Multi-view camera setups.** The acquisition system uses five Microsoft Azure Kinect devices arranged in symmetric horizontal configuration at $0°$, $\pm30°$, and $\pm45°$ angles within a standardized 556cm$\times$274cm indoor environment, as illustrated in Fig. 3. While cameras maintain horizontal symmetry, each device operates at different heights and poses to enhance viewpoint diversity. Camera heights and poses are randomly adjusted every 10 groups to capture more diverse viewpoints.

Participants move freely within marked activity areas and perform each action four times per session to create different viewpoints: facing the cameras, facing away from cameras, switching positions while facing cameras, and switching positions while facing away. This collection protocol results in 20 different camera views for each participant pair performing the same action, ensuring comprehensive coverage from multiple angles, especially challenging rear views that are often missing in existing datasets. For interactive actions, participants also switch their active and passive roles when changing positions to capture both sides of the interaction. To ensure realistic performances while maintaining safety, we use appropriate items for different action types: toy weapons for simulated violence, wigs for hair-pulling actions, tissues for mouth-covering gestures, and soft objects like paper boxes for hitting actions to prevent injury.

**Data prerocessing.** We developed custom software to manage data collection across all five synchronized Azure Kinect cameras. The software records the exact start and end time of each action, ensuring all cameras capture the same actions simultaneously. During data processing, we use these recorded timestamps to extract individual action clips from the complete recordings of each group. Each clip contains three types of data: RGB, depth, and 3D skeleton videos, as shown in Fig. 1. For actions that are naturally short, we extend them to the standard 300-frame length by repeating the action frames.

**Dataset statistics.** ANUBIS comprises 102 action categories collected from 80 participants, generating 66,232 skeleton clips across 80 viewpoints, as presented in Tab.I. The viewpoint distribution includes 40 frontal views and 40 rear views from different angles, ensuring balanced coverage between frontal perspectives and challenging posterior orientations. Based on action categories, the dataset contains 45 independent actions (single-person behaviors) and 57 multi-person interactions. Among multi-person actions, we include 17 social interaction behaviors (e.g., handshaking, patting shoulders, object exchange), and 40 aggressive actions (e.g., hitting, stabbing, strangling). The complete statistics of ANUBIS are shown in Fig.4.

TABLE I: Comprehensive Overview of Skeleton-based Action Recognition Datasets

| Dataset | Classes | Views | Subjects | Clips | Sensors | Additional Modalities | Dataset Type |
|---|---|---|---|---|---|---|---|
| HDM05 (2007) [57] | 130 | 1 | 5 | 2,337 | - | RGB | Human motion capture |
| MSRAction3D (2010) [45] | 20 | 1 | 10 | 567 | Kinect | RGB, Depth | Daily activities |
| CAD-60 (2011) [84] | 12 | - | 4 | 68 | Kinect | RGB, Depth | Human performing activities |
| MSRDailyActivity3D (2012) [89] | 16 | 1 | 10 | 320 | Kinect | RGB, Depth | Daily activities |
| G3D-Gaming (2012) [3] | 20 | 1 | 10 | - | Kinect | RGB, Depth | Gaming gestures |
| UTKinect (2012) [113] | 10 | - | 10 | 200 | Kinect | RGB, Depth | Human actions |
| SBU (2012) [123] | 8 | - | 7 | 282 | Kinect | RGB | Human-human interaction |
| CAD-120 (2013) [35] | 10 | - | 4 | 120 | Kinect | RGB, Depth | Activity types & object interactions |
| Berkeley MHAD (2013) [60] | 11 | 4 | 12 | 660 | Kinect | RGB, Depth, Audio, Accelerometer | Multimodal Capture & Controllable & Synced Data |
| Florence3D-Action (2013) [72] | 9 | - | 10 | 215 | Kinect | RGB, Depth | Daily Activities |
| MSRActionPairs3D (2013) [61] | 12 | - | 10 | 360 | Kinect | RGB, Depth | 3D Action & Gesture Recognition |
| UCFKinect (2013) [23] | 16 | - | 16 | 1,280 | - | RGB, Depth | General actions |
| Northwestern-UCLA (2014) [90] | 10 | 3 | 10 | 1,494 | Kinect | RGB, Depth | Daily Activities |
| Multi-View TJU (2014) [46] | 20 | 2 | 22 | 7,040 | - | RGB, Depth | Multi-view actions |
| UWA3D Multiview Activity (2014) [70] | 30 | 4 | 10 | 701 | Kinect | RGB, Depth | Multi-view actions |
| SYSU 3D HOI (2015) [25] | 12 | - | 40 | 480 | Kinect | RGB, Depth | Human-object interaction |
| UWA3D Multiview Activity II (2015) [69] | 30 | 4 | 10 | 1,070 | Kinect | RGB, Depth | Daily activities |
| NTU-60 (2016) [73] | 60 | 80 | 40 | 56,880 | Kinect v2 | RGB, Depth, Infrared | Large-scale general actions |
| PKU-MMD I (2017) [47] | 51 | 3 | 66 | 1,076 | Kinect v2 | RGB, Depth, Infrared | Multi-modal actions |
| Kinetics-skeleton (2018) [119] | 400 | - | - | 260,232 | - | - | Based on publicly available RGB videos |
| RGB-D Varying-View (2018) [30] | 40 | 9 | 118 | 25,600 | Kinect v2 | RGB, Depth | Multi-view actions |
| NTU-120 (2019) [51] | 120 | 155 | 106 | 114,480 | Kinect v2 | RGB, Depth, Infrared | Large-scale general actions |
| MMAct (2019) [33] | 37 | 5 | 20 | 36,764 | - | RGB, Accelerometer, Gyroscope | Multi-modal actions |
| PKU-MMD II (2020) [50] | 41 | 3 | 13 | 1,009 | Kinect v2 | RGB, Depth, Infrared | Multi-modal actions |
| ETRI-Activity3D (2020) [29] | 55 | - | 100 | 112,620 | Kinect v2 | RGB | Daily activities of the elderly |
| IKEA ASM (2020) [2] | 33 | 3 | 48 | 16,764 | Kinect v2 | RGB, Depth | Furniture assembly |
| UAV-Human (2021) [44] | 155 | - | 119 | 22,476 | Azure Kinect | RGB, Infrared, Depth | UAV perspective actions |
| NCRC (2022) [28] | 6 | - | 8 | 398 | - | - | Nursing care activities |
| Tai-Chi (2022) [122] | 10 | - | - | 200 | Perception Neuron | - | Martial arts |
| ANUBIS (2025) | 102 | 80 | 80 | 66,232 | Azure Kinect | RGB, Depth | Large-Scale & Multi-Person & Frontal / Rear-View & In-the-Wild |

### B. Comparison with Existing Datasets

Table I presents a comparative summary of our dataset against existing benchmarks.

Existing datasets exhibit diverse characteristics across devices, modalities, and application scenarios. NTU-60 [73] establishes a multi-view 3D action analysis benchmark with 60 indoor action categories, while NTU-120 [51] extends the action repertoire to 120 classes to enhance classification challenges. Meanwhile, Kinetics-skeleton [119] extracts skeletal information from large-scale RGB videos to provide structured representations. For multimodal fusion, PKU-MMD I [47] integrates RGB, depth, and infrared data to support continuous action modeling, whereas PKU-MMD II [50] builds upon this multimodal foundation by adding fine-grained annotations for interaction actions. Additionally, MMAct [33] combines visual and inertial sensor data to accommodate mobile scenarios, and RGB-D Varying-View [30] specifically explores viewpoint robustness through dynamic perspective changes. In vertical applications, ETRI-Activity 3D [29] focuses on elderly daily activity monitoring, while IKEA ASM [2] targets fine-grained manipulation tasks such as furniture assembly. Furthermore, NCRC-Human [28] concentrates on nursing scenario action analysis, and Tai-Chi [122] provides in-depth characterization of Tai Chi movement kinematics. Moreover, UAV-Human [44] enriches spatial observation dimensions through drone perspectives, and several datasets use devices like Azure Kinect DK to improve acquisition precision. These datasets collectively serve different research objectives including general algorithm validation, scenario-specific optimization, and cross-modal learning, forming a diversified data ecosystem for research.

ANUBIS offers distinct advantages over existing datasets in three key aspects. First, in terms of technical specifications, we use Microsoft Azure Kinect devices, providing superior skeleton extraction accuracy and stability compared to Kinect V1/V2 systems used in most previous benchmarks. The dataset captures 32 joint coordinates per skeleton, adding 7 additional joints compared to the 25 joints in NTU datasets: 5 facial joints (nose, eyes, and ears) and 2 clavicle joints, enabling more detailed representation of facial expressions and upper body posture. Second, regarding content scope and diversity, ANUBIS addresses critical gaps in existing benchmarks by expanding action categories. While NTU-60 (60 categories, 40 participants, 56,880 sequences) and NTU-120 (120 categories, 106 participants, 114,480 sequences) established foundational benchmarks emphasizing scale expansion, ANUBIS (102 categories, 80 participants, 66,232 clips) introduces qualitative advances through three previously underrepresented action types: complex multi-person interactions for collaborative behavior analysis, aggressive behaviors essential for security and surveillance applications, and contemporary pandemic-induced social adaptations reflecting evolving interaction patterns. Most importantly, the most distinctive innovation lies in comprehensive viewpoint coverage through systematic rear-view data collection across multiple angles. This creates novel technical challenges where hand details and frontal movements become occluded, better approximating real-world monitoring conditions where subjects may not always face cameras directly. This balanced frontal-posterior perspective distribution addresses critical observational limitations in traditional datasets while enhancing model robustness under diverse camera orientations essential for practical deployment scenarios.

## IV. Evaluation and Benchmarking

### A. Experimental Setups

We benchmarked a range of state-of-the-art skeleton-based action recognition methods on our newly collected ANUBIS dataset and evaluated these methods on the NTU datasets for comparative analysis. All experiments were implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU for 50 epochs. Stochastic Gradient Descent (SGD) with momentum 0.9 was used as the optimizer, with an initial learning rate of 0.05, decayed to 10% at epoch 30.

Skeleton data was preprocessed via normalization and translation [103]. All video clips were standardized to 300 frames using action repetition (except SkateFormer [17], which retained

TABLE II: Performance of representative skeletal action recognition methods on NTU60, NTU120, and the proposed ANUBIS. While NTU60 and NTU120 results are approaching saturation, ANUBIS presents a substantially more challenging and unsolved benchmark, leaving significant room for performance improvement. The **best** and **second-best** performances are highlighted.

| Method | Venue | Features | Dataset | | | | | | Params (M) | GFLOPs |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | NTU60 | | NTU120 | | ANUBIS | | | |
| | | | X-Sub | X-View | X-Sub | X-Set | Top-1 | Top-5 | | |
| STGCN [118] | **AAAI** 2018 | Joint | 81.5 | 88.3 | - | - | 50.25 | 79.96 | 3.4 | 45.23 |
| Motif-STGCN [111] | **AAAI** 2019 | Joint | 84.2 | 90.2 | - | - | 55.76 | 83.96 | 1.78 | 27.10 |
| 2s-AGCN [76] | **CVPR** 2019 | Joint+Bone | 88.5 | 95.1 | - | - | 57.26 | 84.86 | 3.47 | 47.84 |
| MS-G3D [54] | **CVPR** 2020 | Joint+Bone | 91.5 | 96.2 | 86.9 | 88.4 | 54.17 | 82.05 | 3.8 | 62.72 |
| GCN-NAS [64] | **AAAI** 2020 | Joint+Bone | 89.4 | 95.7 | - | - | 56.40 | 84.37 | 6.57 | 93.64 |
| HD-GCN [40] | **ICCV** 2023 | Joint+Bone | 93.4 | 97.2 | 90.1 | 91.6 | 51.33 | 80.96 | 8.8 | 12.74 |
| RA-GCN [81] | **ICIP** 2019 | Joint+Motion | 85.9 | 93.5 | - | - | 41.87 | 73.45 | 10.26 | 135.52 |
| Shift-GCN [10] | **ICIP** 2019 | Joint+Bone+Motion | 90.7 | 96.5 | 85.9 | 87.6 | 26.84 | 57.50 | 0.73 | 6.16 |
| Decoupling-GCN [9] | **ECCV** 2020 | Joint+Bone+Motion | 90.8 | 96.6 | 86.5 | 88.1 | 52.32 | 80.06 | 3.63 | 32.98 |
| CTR-GCN [8] | **ICCV** 2021 | Joint+Bone+Motion | 92.4 | 96.8 | 88.9 | 90.4 | 37.90 | 70.40 | 10.07 | 141.68 |
| STTFormer [67] | **arxiv** 2022 | Joint+Bone+Motion | 92.3 | 96.5 | 88.3 | 89.2 | 57.77 | 85.30 | 6.6 | 109.08 |
| Hyperformer [129] | **arxiv** 2022 | Joint+Bone+Motion | 92.9 | 96.5 | 89.9 | 91.3 | 47.51 | 77.77 | 3.1 | 32.68 |
| InfoGCN [12] | **CVPR** 2022 | Joint+Bone+Motion | 93.0 | 97.1 | 89.4 | 90.7 | 46.99 | 76.69 | 1.6 | 19.97 |
| BlockGCN [130] | **CVPR** 2024 | Joint+Bone+Motion | 93.1 | 97.0 | 90.3 | 91.5 | 54.46 | 81.73 | 2.5 | 36.79 |
| Skateformer [17] | **ECCV** 2024 | Joint+Bone+Motion | **93.5** | **97.8** | 89.8 | 91.4 | 45.02 | 75.43 | 3.8 | 8.93 |
| DS-STGCN [115] | **TIP** 2024 | Joint+Bone+Motion | 93.2 | **97.5** | 89.4 | 91.2 | 52.43 | 81.96 | 1.4 | 14.09 |
| DeGCN [58] | **TIP** 2024 | Joint+Bone+Motion | **93.6** | 97.4 | **91.0** | **92.1** | **60.16** | **85.63** | 1.4 | 9.75 |
| ProtoGCN [48] | **CVPR** 2025 | Joint+Bone+Motion | **93.5** | **97.5** | 90.4 | **91.9** | 47.56 | 78.10 | 4.2 | 29.88 |
| LA-GCN [116] | **TMM** 2025 | Joint+Bone+Motion | **93.5** | 97.2 | **90.7** | 91.8 | **60.33** | **86.87** | 3.4 | 28.32 |

its original 64-frame window). Evaluation metrics included Top-1 and Top-5 classification accuracy, as well as model complexity indicators. To show the recognition accuracies of a model for all the action classes, a confusion matrix is used [91], [93].

### B. Benchmark Comparison: NTU vs. ANUBIS

Table II presents our benchmark comparison, followed by a detailed analysis of each aspect.

**Performance saturation on NTU.** Results on NTU60 and NTU120 show a clear trend toward performance saturation. Many recent methods, including DeGCN, LA-GCN, and ProtoGCN, exceed 93% Top-1 accuracy on NTU60 cross-subject and over 97% on cross-view splits. On NTU120, these methods regularly surpass 90%, leaving minimal headroom for further improvement. This plateau suggests that NTU datasets, while historically instrumental, no longer fully differentiate the capabilities of state-of-the-art skeleton-based recognition models. Consequently, incremental gains on NTU may not reflect real-world robustness or generalization ability.

**ANUBIS as a more challenging benchmark.** In contrast, ANUBIS results are substantially lower, even for top-performing models. LA-GCN achieves the highest Top-1 accuracy at 60.33%, closely followed by DeGCN at 60.16%, with Top-5 accuracies just above 85%. Most other methods fall in the 40-55% range, despite reaching near-perfect scores on NTU. This performance drop highlights the increased difficulty of ANUBIS, driven by factors such as rear-view occlusions, fine-grained actions, and modern social interaction patterns. The dataset's complexity disrupts conventional represenration exploitation patterns, indicating that current architectures, optimized for standard benchmarks, struggle to adapt to more realistic and diverse action scenarios.

**Representation-type trends and variability.** Interestingly, ANUBIS does not show a straightforward correlation between the number of input feature types and performance. Joint-only (J) methods like Motif-STGCN reach 55.76%, surpassing some tri-representation (Joint + Bone + Motion) methods such as CTR-GCN (37.90%) and Shift-GCN (26.84%). Even within tri-representation designs, variance is high: LA-GCN and DeGCN lead with 60%+ Top-1, while several others hover below 50%. This suggests that the quality of feature fusion strategies and architectural adaptability outweighs simply increasing feature types. On ANUBIS, fusion design, attention mechanisms, and temporal-spatial reasoning appear more critical than the raw presence of multiple features.

**Computational efficiency considerations.** From a resource perspective, there is no consistent trade-off between accuracy and efficiency. Some lightweight models, such as DeGCN (1.4M parameters, 9.75 GFLOPs), achieve top-tier performance, while others with far greater complexity, such as CTR-GCN (10.07M parameters, 141.68 GFLOPs), perform significantly worse on ANUBIS. This efficiency-performance decoupling reinforces that computational cost does not guarantee robustness on challenging datasets. It also underscores the potential for more resource-friendly yet highly accurate architectures, particularly for deployment in edge and real-time scenarios.

## V. In-Depth Analysis and Discussion

### A. Confusion Matrix Analysis and Dataset Challenges

Below, we present an in-depth analysis of confusion matrix for ANUBIS using LA-GCN (best model), shown in Fig. 5.

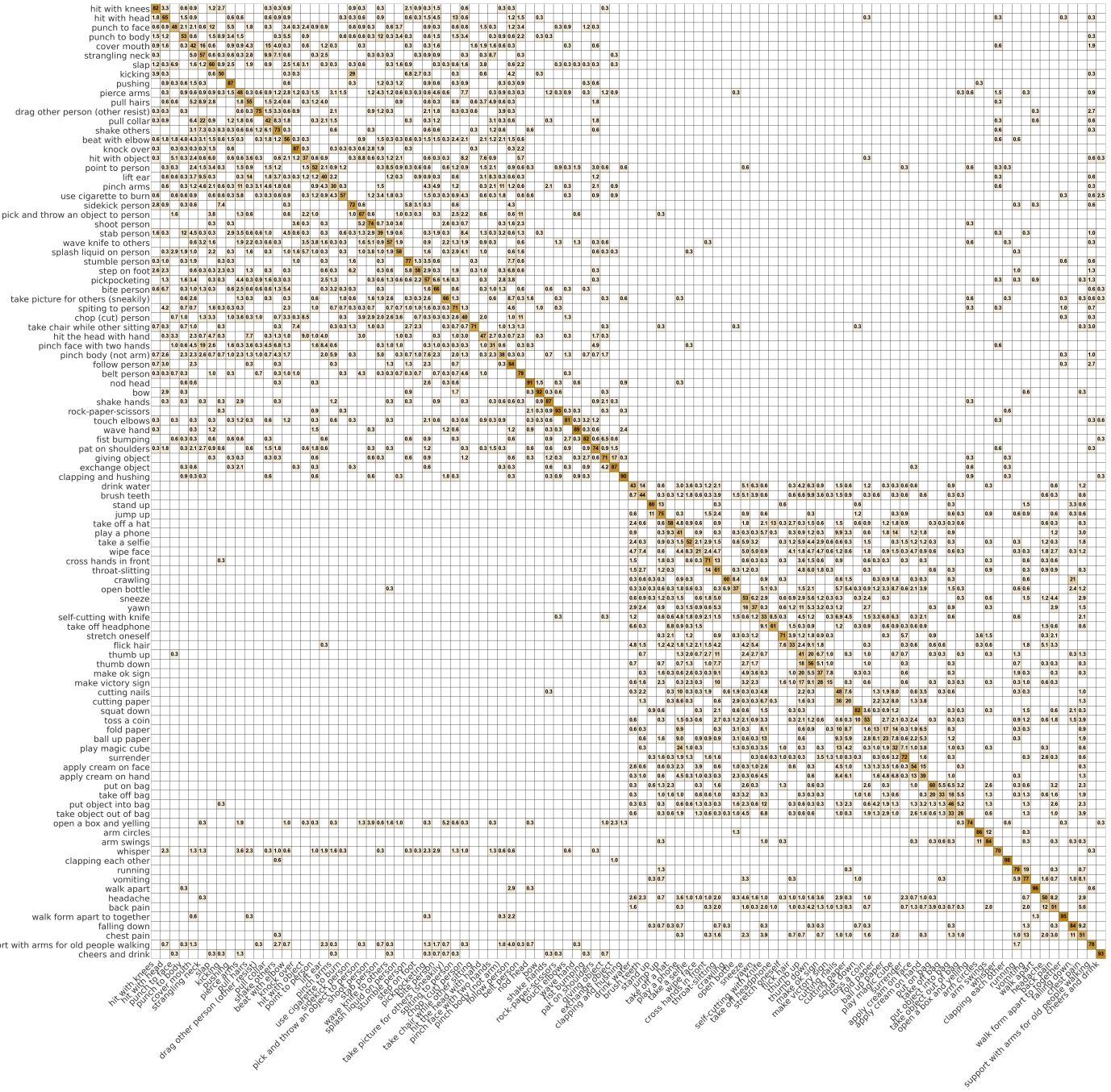



Fig. 5: Confusion matrix of LA-GCN on ANUBIS. Clear diagonal dominance reflects strong recognition for macro-motion and dyadic actions, while off-diagonal confusions cluster among fine-grained, hand/object-centric classes (e.g., fold paper, play magic cube), highlighting the limitations of skeleton-only features in capturing subtle manipulations and object context.

**Overall performance.** The confusion matrix for LA-GCN on ANUBIS shows a highly non-uniform distribution of recognition accuracy across action classes. Certain categories, particularly those with distinctive, large-scale body movements such as clapping each other, walk apart, and walk from apart to together, form bright diagonal clusters, indicating high recognition accuracy with minimal confusion. These actions provide strong spatio-temporal signals and distinctive inter-joint relationships, which LA-GCN can model effectively. In contrast, many fine-grained or localized actions display heavy off-diagonal activity, reflecting substantial misclassification. The imbalance between well-separated macro-movements and ambiguous micro-movements underscores the diverse difficulty spectrum.

**Sources of misclassification.** High confusion rates are particularly evident among actions with similar skeletal motion trajectories but differing in object interactions or subtle hand gestures, for example, fold paper vs. cutting paper, make victory sign vs. thumb up, or apply cream on face vs. wash face. Since ANUBIS is purely skeleton-based, critical cues such as object presence, fine finger articulation, or texture changes are absent, forcing the model to infer action semantics from incomplete spatial-temporal data. Additionally, rear-view sequences exacerbate this

challenge, as many hand and torso movements become partially occluded, reducing the discriminability of similar gestures.

**Viewpoint and interaction complexity.** A notable difficulty arises from ANUBIS's inclusion of rear-view and multi-person interactive scenarios. In rear-view cases, key discriminative joints (hands, face orientation) are less visible or entirely occluded, leading to higher confusion even for otherwise distinct actions. Meanwhile, interactive actions sometimes benefit from additional relational cues (e.g., body proximity, coordinated movements), improving recognition compared to isolated fine-motor actions. However, when multiple participants perform overlapping motions, inter-person occlusions and role ambiguity can also increase error rates, as seen in confusions between certain cooperative and antagonistic interactions.

**Why ANUBIS is a valuable benchmark?** This confusion matrix shows why ANUBIS is substantially more challenging than traditional benchmarks like NTU. It combines viewpoint variation, fine-grained gestures, modern social behaviors, and complex multi-person interactions, all of which stress-test a model's spatial reasoning, temporal modeling, and robustness to incomplete information. For future researchers, ANUBIS offers a fertile ground to explore: (i) Multi-modal fusion: integrating RGB, depth, or object-context streams to resolve semantic ambiguities. (ii) Fine-grained local feature learning: improving finger joint precision, using high-resolution pose estimations, and applying attention mechanisms to hand regions. (iii) Viewpoint-invariant representations: developing architectures resilient to occlusions and perspective shifts. (iv) Relational reasoning: enhancing modeling of inter-person dynamics for both cooperative and competitive interactions.

**Implications for model design.** The persistent confusions highlighted here suggest that future algorithms must go beyond current GCN and Transformer hybrids, which excel in macro-motion capture but falter in local detail extraction. Promising directions include hierarchical feature learning, where models capture both global body structure and localized joint dynamics; self-supervised pretraining to enrich motion semantics without additional labels; and scene- or object-aware skeleton augmentation to inject missing contextual signals. By tackling the challenges exposed by ANUBIS, researchers can develop models with stronger generalization, adaptability to real-world occlusions, and robustness to nuanced human behaviors, capabilities essential for next-generation action understanding systems.

### *B. Action-Level Performance Trends on ANUBIS*

Below we provide a focused, multi-angle discussion of the per-class results (Table III) for the six best models on ANUBIS.

**What easy vs. hard classes reveal?** Across models, the top-20 classes average 84.6% accuracy, while the bottom-20 average only 27.5%. High-accuracy actions (e.g., clapping each other 96-98%, walk apart 96-97% for most models, bow up to 95.1%) share clear, whole-body motion signatures and/or large spatial displacements that produce distinctive spatio-temporal patterns. In contrast, the hardest classes are dominated by fine-motor, hand-centric, or human-object interactions (e.g., fold paper, ball up paper, play magic cube, open bottle), where skeletal streams alone underspecify the semantics (no object context; limited finger fidelity). This split aligns with the broader finding that skeleton-only pipelines excel at macro body dynamics but underperform on micro-manipulation.

**Cross-model agreement as a stability signal.** The Range column (max-min accuracy across models) is a useful consensus proxy. It's smaller for easy classes (mean range $\sim$ 0.105 in the top-20) and larger for hard classes (mean range $\sim$ 0.169 in the bottom-20), indicating that difficult categories induce greater architectural sensitivity. Stable/easy classes include clapping each other (range 0.058), knock over (0.056), rock-paper-scissors (0.050), walk from apart to together (0.047), and nod head (0.044). By contrast, high-variance categories, put object into bag (0.320), open bottle (0.271), flick hair (0.254), play magic cube (0.218), show strong model-specific behavior, suggesting that fusion design and local feature modeling, rather than modality count, drive the differences.

**Model-level profiles and robustness.** Averaged over the top-20, LA-GCN leads (87.9%), followed by STTFormer (86.9%), Motif-STGCN (84.3%), ST-GCN (84.0%), 2s-AGCN (83.3%), and DeGCN (81.2%). On the bottom-20, LA-GCN again ranks first (32.1%), then STTFormer (29.1%), DeGCN (27.7%), Motif-STGCN (26.8%), 2s-AGCN (25.6%), and ST-GCN (23.9%). Notably, DeGCN has the smallest accuracy drop from top-20 to bottom-20 ($\sim$ 53.5 pp vs. 55-60 pp for others), indicating relatively better resilience on difficult, fine-grained classes, even though its absolute accuracy remains lower on the easy set. This pattern hints that DeGCN's part/decoupling bias helps when global motion cues are weak.

**Who wins where (per-class wins)?** Counting ties as shared wins: in the top-20, LA-GCN leads with 8 wins (e.g., walk from apart to together, rock-paper-scissors, clapping and hushing), STTFormer follows with 6 (e.g., bow, knock over, squat down), and Motif-STGCN contributes 4 (e.g., walk apart, falling down, touch elbows, whisper). In the bottom-20, LA-GCN again leads (9 wins), while DeGCN notably secures 5 wins (e.g., open bottle, make victory sign, ball up paper, fold paper), reinforcing its relative strength on hand/object-centric actions. These distributions suggest LA-GCN offers the best overall balance, while DeGCN is disproportionately helpful on the tail where local articulation dominates.

**Action taxonomy: dyadic vs. single-subject fine motor.** Many top classes are dyadic or coordinated (clapping each other, exchange object, shake hands, pushing, follow person, whisper). The presence of a partner supplies relative pose and motion cues that are easier to encode in graphs/transformers, boosting separability. Conversely, bottom classes are mostly single-subject, object-centric, or subtle gestures (fold paper, ball up paper, play magic cube, make ok sign, yawn). Here, object state and finger articulation, both weakly represented in standard skeletons, are decisive. The outcome strongly argues for object-aware and hand-aware augmentations (e.g., explicit hand-pose subgraphs; contact tokens; object proxy nodes inferred from motion; or lightweight RGB/IR cues fused selectively).

**Where each architecture fails and why?** Architectural biases surface in outliers. For instance, DeGCN underperforms on walk apart (83.8%) relative to others ($\sim$96-97%), suggesting that some decoupled designs may under-leverage global displacement and long-range cross-person cues. STTFormer shows excellent performance on macro patterns (bow 95.1%) but can collapse on very fine manipulation (fold paper 2.5%), hinting

TABLE III: Per-class accuracy of six leading models on ANUBIS, showing the top-20 easiest and bottom-20 most challenging actions. Easy classes are dominated by large-scale body motions and dyadic interactions, yielding high consensus across models, while hard classes involve fine-grained, hand/object-centric gestures with high variance, showing architectural sensitivity and representation limitations.

| Rank | Action Name | LA-GCN | STTFormer | 2s-AGCN | STGCN | Motif-STGCN | DeGCN | Range |
|---|---|---|---|---|---|---|---|---|
| 1 | clapping each other | 0.984 | 0.984 | 0.926 | 0.955 | 0.952 | 0.964 | 0.058 |
| 2 | walk apart | 0.958 | 0.971 | 0.964 | 0.961 | 0.974 | 0.838 | 0.136 |
| 3 | bow | 0.920 | 0.951 | 0.880 | 0.926 | 0.914 | 0.883 | 0.071 |
| 4 | walk form apart to together | 0.950 | 0.934 | 0.934 | 0.903 | 0.903 | 0.906 | 0.047 |
| 5 | cheers and drink | 0.935 | 0.945 | 0.792 | 0.831 | 0.798 | 0.769 | 0.176 |
| 6 | rock-paper-scissors | 0.935 | 0.899 | 0.885 | 0.905 | 0.908 | 0.885 | 0.050 |
| 7 | nod head | 0.912 | 0.898 | 0.877 | 0.921 | 0.918 | 0.886 | 0.044 |
| 8 | arm circles | 0.864 | 0.861 | 0.874 | 0.848 | 0.770 | 0.906 | 0.136 |
| 9 | clapping and hushing | 0.902 | 0.887 | 0.830 | 0.869 | 0.851 | 0.842 | 0.072 |
| 10 | wave hand | 0.888 | 0.885 | 0.831 | 0.799 | 0.814 | 0.793 | 0.089 |
| 11 | knock over | 0.867 | 0.883 | 0.877 | 0.852 | 0.827 | 0.855 | 0.056 |
| 12 | exchange object | 0.874 | 0.828 | 0.720 | 0.735 | 0.774 | 0.780 | 0.154 |
| 13 | shake hands | 0.865 | 0.815 | 0.871 | 0.818 | 0.789 | 0.733 | 0.138 |
| 14 | pushing | 0.870 | 0.818 | 0.815 | 0.809 | 0.762 | 0.738 | 0.132 |
| 15 | arm swings | 0.845 | 0.861 | 0.786 | 0.809 | 0.854 | 0.764 | 0.097 |
| 16 | falling down | 0.839 | 0.803 | 0.822 | 0.783 | 0.855 | 0.809 | 0.072 |
| 17 | touch elbows | 0.811 | 0.814 | 0.702 | 0.794 | 0.847 | 0.770 | 0.145 |
| 18 | squat down | 0.816 | 0.842 | 0.750 | 0.777 | 0.753 | 0.711 | 0.131 |
| 19 | follow person | 0.840 | 0.756 | 0.786 | 0.686 | 0.766 | 0.706 | 0.154 |
| 20 | whisper | 0.702 | 0.735 | 0.744 | 0.812 | 0.832 | 0.709 | 0.130 |
| 83 | put object into bag | 0.456 | 0.330 | 0.136 | 0.311 | 0.204 | 0.311 | 0.320 |
| 84 | flick hair | 0.329 | 0.456 | 0.381 | 0.202 | 0.423 | 0.456 | 0.254 |
| 85 | hit with object | 0.366 | 0.423 | 0.423 | 0.390 | 0.441 | 0.390 | 0.075 |
| 86 | pinch body (not arm) | 0.380 | 0.347 | 0.343 | 0.432 | 0.340 | 0.271 | 0.161 |
| 87 | drink water | 0.432 | 0.384 | 0.345 | 0.384 | 0.321 | 0.348 | 0.111 |
| 88 | pinch face with two hands | 0.305 | 0.338 | 0.244 | 0.296 | 0.422 | 0.292 | 0.178 |
| 89 | stab person | 0.390 | 0.348 | 0.413 | 0.319 | 0.381 | 0.332 | 0.094 |
| 90 | thumb up | 0.407 | 0.283 | 0.360 | 0.350 | 0.259 | 0.253 | 0.154 |
| 91 | chop (cut) person | 0.401 | 0.401 | 0.212 | 0.293 | 0.323 | 0.371 | 0.189 |
| 92 | open bottle | 0.374 | 0.229 | 0.262 | 0.121 | 0.241 | 0.392 | 0.271 |
| 93 | make ok sign | 0.367 | 0.286 | 0.195 | 0.247 | 0.166 | 0.211 | 0.201 |
| 94 | yawn | 0.372 | 0.316 | 0.322 | 0.357 | 0.304 | 0.201 | 0.171 |
| 95 | make victory sign | 0.146 | 0.188 | 0.153 | 0.162 | 0.156 | 0.341 | 0.195 |
| 96 | self-cutting with knife | 0.332 | 0.302 | 0.317 | 0.215 | 0.323 | 0.278 | 0.117 |
| 97 | cutting paper | 0.204 | 0.331 | 0.175 | 0.178 | 0.194 | 0.226 | 0.156 |
| 98 | play magic cube | 0.324 | 0.228 | 0.147 | 0.106 | 0.135 | 0.128 | 0.218 |
| 99 | wipe face | 0.209 | 0.239 | 0.227 | 0.195 | 0.268 | 0.198 | 0.073 |
| 100 | take object out of bag | 0.256 | 0.233 | 0.188 | 0.117 | 0.210 | 0.146 | 0.139 |
| 101 | ball up paper | 0.233 | 0.140 | 0.137 | 0.078 | 0.214 | 0.242 | 0.164 |
| 102 | fold paper | 0.134 | 0.025 | 0.134 | 0.019 | 0.040 | 0.158 | 0.139 |

at insufficient high-precision local attention or challenges with ambiguous rear-view hand cues. Motif-STGCN, despite being older, excels on global, rhythmically coherent actions (walk apart 97.4%, falling down 85.5%, whisper 83.2%), likely benefiting from recurring motion motifs. These contrasts imply that ensembling or hybridizing (e.g., LA-GCN backbone + DeGCN-style local decoupling + motif priors) could yield gains.

**Practical guidance and future directions.** For general deployment, LA-GCN is the most reliable head-and-tail performer. For hand/object-heavy applications, consider mixing a DeGCN-like local articulation module or hand-focused subgraphs. Across the board, the bottom classes make a strong case for: (i) Multi-modal enrichment: add lightweight object/context cues (RGB patches, object heatmaps) or learned object nodes linked to hands. (ii) Local attention: high-resolution, hand-centric attention (hierarchical GCNs/Transformers, dilated temporal windows) to capture micro-gestures. (iii) Interaction modeling: explicit cross-person relational edges and contact events for dyadic actions. (iv) Curriculum/augmentation: viewpoint-hard negatives, hand-pose perturbations, and synthetic object-interaction variations to reduce overfitting to macro motion.

Beyond Top-1/Top-5, track (i) per-class stability via Range, (ii) hand/object subset scores (tail classes), and (iii) dyadic vs. single-subject splits. Reporting these alongside overall accuracy makes progress on ANUBIS more diagnostic and reduces the risk of gains being driven by already-easy, macro-motion categories. ANUBIS is a valuable resource for future research.

### C. Analysis of Feature Type Impact

Below, we analyze the impact of feature types on the top three performing models on our ANUBIS benchmark.

**Performance variation across actions.** Fig. 6 presents the per-class accuracy distribution for the top-performing models on ANUBIS, offering a detailed view of how recognition perfor-

TABLE IV: Actions with consistent accuracy gains across all three models (LA-GCN, STTFormer, DeGCN) when adding the Bone feature type. Only 7 out of 102 classes show universal benefit, indicating that bone vectors selectively help actions where limb geometry and joint relationships are key (e.g., arm-hand positioning).

| Rank | Action Name | LA-GCN | | | STTFormer | | | DeGCN | | | Avg. Improve |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Joint | Joint+Bone | Improve | Joint | Joint+Bone | Improve | Joint | Joint+Bone | Improve | |
| 1 | sneeze | 0.5310 | 0.5693 | +0.0383 | 0.3628 | 0.5192 | +0.1564 | 0.4779 | 0.5310 | +0.0531 | **+0.0826** |
| 2 | thumb up | 0.4074 | 0.4175 | +0.0101 | 0.2828 | 0.4141 | +0.1313 | 0.2525 | 0.3468 | +0.0943 | **+0.0786** |
| 3 | follow person | 0.8395 | 0.8696 | +0.0301 | 0.7559 | 0.8462 | +0.0903 | 0.7057 | 0.7559 | +0.0502 | **+0.0569** |
| 4 | running | 0.7883 | 0.7948 | +0.0065 | 0.7362 | 0.8241 | +0.0879 | 0.7557 | 0.8013 | +0.0456 | **+0.0467** |
| 5 | pull collar | 0.4202 | 0.4448 | +0.0246 | 0.3834 | 0.4141 | +0.0307 | 0.3006 | 0.3804 | +0.0798 | **+0.0450** |
| 6 | self-cutting with knife | 0.3323 | 0.3867 | +0.0544 | 0.3021 | 0.3263 | +0.0242 | 0.2779 | 0.2870 | +0.0091 | **+0.0292** |
| 7 | punch to face | 0.4787 | 0.5488 | +0.0701 | 0.4909 | 0.4939 | +0.0030 | 0.3994 | 0.4116 | +0.0122 | **+0.0284** |

TABLE V: Actions improved in at least two of three models when adding the Motion feature type. No action achieved consistent gains across all models, underscoring the model-dependent and unstable nature of motion integration, beneficial for certain temporally distinctive actions but harmful in others. Action abbreviations: "support old people walking" refers to "support with arms for old people walking".

| Rank | Action Name | LA-GCN | | | STTFormer | | | DeGCN | | | Avg. Improve |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Joint | Joint+Motion | Improve | Joint | Joint+Motion | Improve | Joint | Joint+Motion | Improve | |
| 1 | stand up | 0.7964 | 0.7725 | -0.0239 | 0.4820 | 0.6617 | +0.1797 | 0.5719 | 0.7605 | +0.1886 | **+0.1148** |
| 2 | play magic cube | 0.3237 | 0.2596 | -0.0641 | 0.2276 | 0.3558 | +0.1282 | 0.1282 | 0.3750 | +0.2468 | **+0.1036** |
| 3 | take off a hat | 0.5761 | 0.7104 | +0.1343 | 0.5493 | 0.5284 | -0.0209 | 0.3851 | 0.4567 | +0.0716 | **+0.0617** |
| 4 | play a phone | 0.4149 | 0.5403 | +0.1254 | 0.4239 | 0.3403 | -0.0836 | 0.2657 | 0.3791 | +0.1134 | **+0.0517** |
| 5 | cutting paper | 0.2038 | 0.3726 | +0.1688 | 0.3312 | 0.3121 | -0.0191 | 0.2261 | 0.2261 | +0.0000 | **+0.0499** |
| 6 | running | 0.7883 | 0.8241 | +0.0358 | 0.7362 | 0.8143 | +0.0781 | 0.7557 | 0.7362 | -0.0195 | **+0.0315** |
| 7 | support old people walking | 0.7800 | 0.6967 | -0.0833 | 0.6500 | 0.7067 | +0.0567 | 0.5700 | 0.6700 | +0.1000 | **+0.0245** |
| 8 | squat down | 0.8155 | 0.8452 | +0.0297 | 0.8423 | 0.7887 | -0.0536 | 0.7113 | 0.8065 | +0.0952 | **+0.0238** |
| 9 | jump up | 0.7545 | 0.7455 | -0.0090 | 0.6108 | 0.7156 | +0.1048 | 0.7126 | 0.6826 | -0.0300 | **+0.0219** |
| 10 | pull collar | 0.4202 | 0.3957 | -0.0245 | 0.3834 | 0.4325 | +0.0491 | 0.3006 | 0.3712 | +0.0706 | **+0.0317** |

TABLE VI: Actions with consistent accuracy drops across all three models when adding either Bone or Motion features (worst-affected feature type reported). All top declines are linked to Motion, with some drops exceeding 40%, highlighting the risk of unfiltered motion cues overwhelming stable joint-based representations. Action abbreviations: "walk apart together" refers to "walk form apart to together" and "throw object to person" refers to "pick and throw an object to person".

| Rank | Action Name | Feature | LA-GCN | | | STTFormer | | | DeGCN | | | Avg. Decline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Joint | Added feature | Decline | Joint | Added feature | Decline | Joint | Added feature | Decline | |
| 1 | walk apart together | Motion | 0.9497 | 0.8365 | -0.1132 | 0.9340 | 0.2799 | -0.6541 | 0.9057 | 0.3491 | -0.5566 | **-0.4413** |
| 2 | walk apart | Motion | 0.9579 | 0.3042 | -0.6537 | 0.9709 | 0.8123 | -0.1586 | 0.8382 | 0.5761 | -0.2621 | **-0.3581** |
| 3 | surrender | Motion | 0.7212 | 0.6154 | -0.1058 | 0.7308 | 0.4872 | -0.2436 | 0.7372 | 0.4423 | -0.2949 | **-0.2148** |
| 4 | bite person | Motion | 0.6592 | 0.5732 | -0.0860 | 0.6369 | 0.3949 | -0.2420 | 0.6561 | 0.3631 | -0.2930 | **-0.2070** |
| 5 | fist bumping | Motion | 0.8190 | 0.6499 | -0.1691 | 0.7774 | 0.5905 | -0.1869 | 0.5816 | 0.3917 | -0.1899 | **-0.1820** |
| 6 | back pain | Motion | 0.5131 | 0.4739 | -0.0392 | 0.6176 | 0.4216 | -0.1960 | 0.5817 | 0.2745 | -0.3072 | **-0.1808** |
| 7 | throw object to person | Motion | 0.6730 | 0.5143 | -0.1587 | 0.7143 | 0.5143 | -0.2000 | 0.5556 | 0.3778 | -0.1778 | **-0.1788** |
| 8 | open bottle | Motion | 0.3735 | 0.2018 | -0.1717 | 0.2289 | 0.1506 | -0.0783 | 0.3916 | 0.1325 | -0.2591 | **-0.1697** |
| 9 | thumb down | Motion | 0.5623 | 0.4815 | -0.0808 | 0.6364 | 0.3939 | -0.2425 | 0.5320 | 0.3906 | -0.1414 | **-0.1549** |
| 10 | strangling neck | Motion | 0.5666 | 0.3746 | -0.1920 | 0.4799 | 0.3715 | -0.1084 | 0.4458 | 0.2817 | -0.1641 | **-0.1548** |

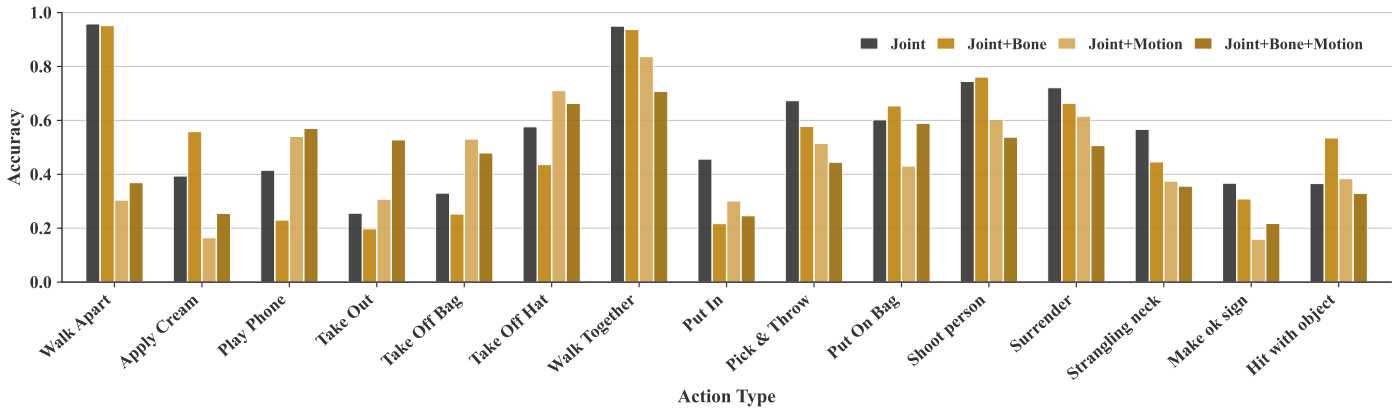



Fig. 6: Analysis of Joint, Bone, Motion data stream effects on action recognition using LA-GCN. Results from 15 actions with significant recognition accuracy fluctuations are shown. Action abbreviations: Apply Cream (apply cream on hand), Take Out (take object out of bag), Walk Together (walk form apart to together), Put In (put object into bag), Pick & Throw (pick and throw an object to person), Support Walk (support with arms for old people walking).

mance varies across different action categories. A clear pattern emerges: actions characterized by large spatial displacement or distinctive global body movement patterns (e.g., walk apart, clapping each other, arm swings) consistently achieve high accuracy, often exceeding 90%. These actions generate strong, coherent spatio-temporal signatures that are easily captured by GCN and Transformer-based architectures.

In contrast, fine-grained actions involving subtle local motions, particularly those dominated by hand and finger movements, tend to yield significantly lower accuracy. Examples include fold paper, make victory sign, and cutting paper, where key discriminative cues are concentrated in a few distal joints. Limitations in current skeleton extraction methods, especially in reliably capturing finger joint trajectories, compound this challenge, leading to frequent misclassifications. Moreover, semantic ambiguity between certain action pairs with similar skeletal trajectories (e.g., different object manipulation tasks) further degrades model performance.

Another noteworthy observation is that dyadic interactive actions generally outperform single-subject fine motor tasks, even when the latter involve relatively simple body movements. The relative positioning, coordinated timing, and interaction cues in multi-person actions provide additional discriminative information that helps models distinguish between similar patterns. This reinforces the need for multi-person relational modeling in future architectures.

**Selective benefits of bone feature integration.** Table IV shows that adding the Bone feature type (joint-to-bone vectors) produces consistent performance gains for only 7 out of 102 actions across the three models (LA-GCN, STTFormer, DeGCN). This finding highlights that Bone features are highly task-selective, benefitting actions where limb geometry, orientation, and relational joint structure are critical for recognition. For example, sneeze and thumb up show average gains of +8.26% and +7.86%, respectively, both actions where subtle arm-hand positional cues are decisive. Interestingly, high-mobility actions like running and follow person also see moderate gains, suggesting that bone vectors may help encode consistent gait and body configuration patterns. However, the overall rarity of universal improvement (only $\sim$7% of the action set) indicates that Bone feature integration is not a universally reliable enhancement and should be applied selectively, perhaps dynamically activated based on action category or scene context.

**Mixed and model-dependent effects of motion feature integration.** In stark contrast, Table V demonstrates that Motion features (velocity/acceleration) yield no actions with unanimous gains across all models. Instead, improvements are model-dependent, with several cases showing substantial boosts for certain architectures while harming others. For instance, play magic cube gains a remarkable +24.68% for DeGCN and +12.82% for STTFormer, yet LA-GCN's performance drops by -6.41%. Similarly, stand up benefits from motion cues in STTFormer and DeGCN (both $>$+17%) but slightly declines in LA-GCN (-2.39%). This inconsistency likely stems from differences in temporal modeling and feature fusion strategies, some architectures exploit motion magnitude effectively, while others are disrupted by noise and irrelevant motion signals. Fine-motor actions like cutting paper also show selective benefit, supporting the idea that motion features may better serve actions with distinct temporal rhythms, but require careful integration to avoid destabilizing spatial representations.

**Identifying actions harmed by additional feature types** Table VI lists the 10 actions most negatively impacted by adding Bone or Motion features, with Motion emerging as the primary culprit in all cases. The degradation can be severe: walk apart together and walk apart suffer average drops of -44.13% and -35.81%, respectively, with some models collapsing from $>$95% accuracy to $<$35%. This suggests that for simple locomotion or static postural actions, motion signals may inject confounding noise, especially when natural variations in speed, occlusion, or skeletal jitter mimic other movement patterns. Complex interactive actions such as surrender and bite person also degrade, implying that motion cues alone may mislead the network when the core discriminative features are spatial configurations or object interactions, not velocity patterns. The consistent harm across all three models indicates a systematic vulnerability in current feature fusion pipelines, where unfiltered or poorly weighted motion signals can overwhelm more stable joint-based representations.

**Implications for feature fusion strategies.** These findings paint a nuanced picture: Bone features are occasionally beneficial but mostly neutral, while Motion features are high-risk, high-reward, capable of substantial gains in certain scenarios but catastrophic losses in others. This calls for adaptive, category-aware feature fusion strategies that can modulate or gate feature type contributions based on action type, scene context, or intermediate model confidence. For example, motion integration might be prioritised for actions with pronounced temporal dynamics (jump up, play magic cube), while deactivated for stable-pose actions (walk apart). Similarly, bone features could be selectively used for fine-hand gestures or limb-orientation-dependent actions. Future architectures could incorporate attention-based feature weighting or meta-learning frameworks to dynamically decide which feature types to emphasise per instance.

### *D. Future Research Directions*

Our in-depth evaluation of ANUBIS highlights persistent limitations in current skeleton-based action recognition approaches, particularly in how they handle heterogeneous action types, feature integration, and real-world deployment. These limitations show opportunities for fundamental advances.

**Adaptive feature-type fusion.** Our results show that the utility of different feature types, joint coordinates, bone vectors, and motion cues, varies dramatically between action categories. For example, large-scale displacement actions like walk apart achieve near-saturation accuracy with joint alone, whereas fine-grained manipulations such as cutting paper depend far more on high-frequency motion cues. However, most existing fusion pipelines treat all features as equally relevant, which can dilute useful signals and even reduce performance.

Future work should replace static fusion with dynamic feature-type selection. Models could learn action-feature relevance mappings via attention gates or controller networks, activating only the most informative features per action instance or even per temporal phase (e.g., preparation, execution, recovery). Beyond selection, fusion should respect semantic hierarchy:

joint often encodes high-level pose, bone captures execution strategies, and motion reflects low-level motion dynamics. Explicitly modeling this hierarchy, potentially with language-aligned semantic anchors, could enable richer cross-feature reasoning and improve both interpretability and generalization.

**Large language model-driven action understanding.** The structured nature of skeleton data makes it an ideal partner for multimodal large language models (MLLMs). While models like CLIP excel in image-text alignment, their temporal reasoning remains weak. The ANUBIS dataset, rich in multi-view, socially interactive, and modern behavior patterns, offers fertile ground for skeleton-informed temporal encoders that strengthen video-language alignment. Key directions include skeleton-text contrastive learning for fine-grained action semantics, skeleton-guided attention mechanisms within MLLMs, and tri-modal fusion of skeleton, video, and text. Skeletons bring unique advantages: robustness to background/lighting, compact representation for long sequences, and built-in privacy preservation. These traits make them valuable for domains such as surveillance, instructional content generation, and accessibility tools. Notably, ANUBIS's rear-view and complex dyadic interactions could help train models that understand subtle interpersonal cues often missed in RGB-based training.

**Social safety and behavioral dynamics analysis.** ANUBIS's diverse interpersonal scenarios create new opportunities for safety-critical applications like bullying or harassment detection. Skeleton-based systems can quantify spatial dynamics, role asymmetries, and temporal escalation patterns without exposing identifiable visual information. This enables early detection of harmful behaviors while preserving privacy.

Further, the dataset's coverage of pandemic-era social norms (e.g., elbow touches, distancing gestures) enables studying cultural and temporal shifts in interaction patterns. Combining skeletal kinematics with interaction graphs could yield models capable of detecting subtle power imbalances, changes in group cohesion, or shifts in emotional state, all valuable for workplace monitoring, school safety, and public health.

**Healthcare, rehabilitation, and cognitive monitoring.** Skeleton-based recognition offers an objective, non-invasive framework for medical and wellness applications. In rehabilitation, precise joint tracking enables quantitative movement quality assessment, replacing subjective clinician scoring. Tailored systems could detect Parkinson's tremors, track stroke recovery symmetry, or adapt training regimens dynamically. In eldercare, gait and balance analysis can feed predictive models for fall prevention, while continuous monitoring of daily activity patterns supports independent living assessments. Beyond physical health, subtle deviations in movement rhythm, coordination, or social behavior may serve as early markers for cognitive decline or mental health issues. The social interaction data in ANUBIS could be leveraged to build behavioral baselines for early intervention in schools, workplaces, and care facilities.

## VI. Conclusion

This work provides a representation-centric review of skeleton-based action recognition, introduces the ANUBIS dataset as a challenging new benchmark, and delivers a thorough evaluation of state-of-the-art models across joint, bone, and motion feature types. Our analysis shows that multi-representational fusion yields highly variable outcomes: bone features are most effective for capturing fine-grained geometric relations and coordinated movements, motion features benefit actions with cyclic patterns or distinct phase transitions, yet for large-scale displacements, synchronized multi-person activities, or static postures, their inclusion can degrade accuracy, in some cases severely. We trace this to the structural heterogeneity of the underlying feature manifolds, Euclidean for joints, Lie group for bones, and tangent space for motion, where naïve fusion introduces redundancy, noise, and semantic mismatch. These findings challenge the prevailing assumption that combining more features inherently improves performance, highlighting the need for task-aware, semantically aligned, and structurally compatible fusion strategies. By offering a dataset rich in back-view perspectives, violent interactions, and multi-person dynamics, ANUBIS not only exposes the limitations of current approaches but also provides a robust foundation for developing the next generation of adaptive, generalizable, and context-aware skeleton-based action recognition models.

## Acknowledgments

We thank Tom Gedeon and Saeed Anwar for providing the devices and funding support for this project.

## References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13778–13790, June 2023.

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021.

[3] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer society conference on computer vision and pattern recognition workshops*, pages 7–12. IEEE, 2012.

[4] Anargyros Chatzitofis, Georgios Albanis, Nikolaos Zioulis, and Spyridon Thermos. A low-cost & real-time motion capture system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21453–21458, June 2022.

[5] Huilin Chen, Lei Wang, Yifan Chen, Tom Gedeon, and Piotr Koniusz. When spatial meets temporal in action recognition. *arXiv preprint arXiv:2411.15284*, 2024.

[6] Qixiang Chen, Lei Wang, Piotr Koniusz, and Tom Gedeon. Motion meets attention: Video motion prompts. In *Asian Conference on Machine Learning*, pages 591–606. PMLR, 2025.

[7] Yifei Chen, Dapeng Chen, Ruijin Liu, Hao Li, and Wei Peng. Video action recognition with attentive semantic units. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10170–10180, October 2023.

[8] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021.

[9] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 536–553. Springer, 2020.

[10] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020.

[11] Ke Cheng, Yifan Zhang, Xiangyu He, Jian Cheng, and Hanqing Lu. Extremely lightweight skeleton-based action recognition with shiftgcn++. *IEEE Transactions on Image Processing*, 30:7333–7348, 2021.

[12] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20186–20196, 2022.

[13] Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz. Learnable expansion of graph operators for multi-modal feature fusion. In *The Thirteenth International Conference on Learning Representations*.

[14] Xi Ding and Lei Wang. Quo vadis, anomaly detection? llms and vlms in the spotlight. *arXiv preprint arXiv:2412.18298*, 2024.

[15] Xi Ding and Lei Wang. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1855–1868, 2025.

[16] Xi Ding and Lei Wang. The journey of action recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1869–1884, 2025.

[17] Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, pages 401–420. Springer, 2025.

[18] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 579–583. IEEE, 2015.

[19] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[20] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, 2022.

[21] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. Skeletr: Towards skeleton-based action recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13634–13644, 2023.

[22] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022.

[23] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J Laviola Jr, and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3):420–436, 2013.

[24] Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph neural network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2263–2275, 2021.

[25] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[26] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[27] Zhen Huang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In *International Conference on Multimedia (MM)*, pages 2122–2130, 2020.

[28] Momal Ijaz, Renato Diaz, and Chen Chen. Multimodal transformer for nursing activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2065–2074, 2022.

[29] Jinhyeok Jang, Dohyung Kim, Cheonshu Park, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10990–10997. IEEE, 2020.

[30] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *Proceedings of the 26th ACM international Conference on Multimedia*, pages 1510–1518, 2018.

[31] Jin-Gong Jia, Yuan-Feng Zhou, Xing-Wei Hao, Feng Li, Christian Desrosiers, and Cai-Ming Zhang. Two-stream temporal convolutional networks for skeleton-based human action recognition. *Journal of Computer Science and Technology*, 35(3):538–550, 2020.

[32] Lipeng Ke, Kuan-Chuan Peng, and Siwei Lyu. Towards to-at spatio-temporal focus for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1131–1139, 2022.

[33] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019.

[34] Piotr Koniusz, Lei Wang, and Anoop Cherian. Tensor representations for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):648–665, 2021.

[35] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013.

[36] Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes. 2024.

[37] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017.

[38] Jungho Lee, Minhyeok Lee, Suhwan Cho, Sungmin Woo, Sungjun Jang, and Sangyoun Lee. Leveraging spatio-temporal dependency for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10255–10264, 2023.

[39] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10444–10453, 2023.

[40] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun 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.

[41] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[42] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3595–3603, 2019.

[43] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2020.

[44] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021.

[45] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010.

[46] An-An Liu, Yu-Ting Su, Ping-Ping Jia, Zan Gao, Tong Hao, and Zhao-Xuan Yang. Multiple/single-view human action recognition via part-induced multitask structural learning. *IEEE transactions on cybernetics*, 45(6):1194–1208, 2014.

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

[48] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. *arXiv preprint arXiv:2411.18941*, 2024.

[49] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29248–29257, 2025.

[50] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–24, 2020.

[51] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Conference on Computer Vision and Pattern Recognition (T-PAMI)*, 2019.

[52] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[53] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In

*Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1656, 2017.

[54] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

[55] L. Minh Dang, Kyungbok Min, Hanxiang Wang, Md. Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[56] Saemi Moon, Myeonghyeon Kim, Zhenyue Qin, Yang Liu, and Dongwoo Kim. Anonymization for skeleton action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15028–15036, 2023.

[57] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Mocap database hdm05. *Institut für Informatik II, Universität Bonn*, 2(7), 2007.

[58] Woomin Myung, Nan Su, Jing-Hao Xue, and Guijin Wang. Degcn: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33:2477–2490, 2024.

[59] Nadhira Noor, Fabianaugie Jametoni, Jinbeom Kim, Hyunsu Hong, and In Kyu Park. Efficient skeleton-based action recognition for real-time embedded systems. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5889–5897, 2024.

[60] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE workshop on applications of computer vision (WACV)*, pages 53–60. IEEE, 2013.

[61] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 716–723, 2013.

[62] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *European Conference on Computer Vision*, pages 605–622. Springer, 2022.

[63] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[64] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *The AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 2669–2676, 2020.

[65] Zhenyue Qin, Pan Ji, Dongwoo Kim, Yang Liu, Saeed Anwar, and Tom Gedeon. Strengthening skeletal action recognizers via leveraging temporal patterns. In *European Conference on Computer Vision*, pages 577–593. Springer, 2022.

[66] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):4783–4797, 2022.

[67] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*, 2022.

[68] Jushang Qiu and Lei Wang. Evolving skeletons: Motion dynamics in action recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1916–1937, 2025.

[69] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016.

[70] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014.

[71] Abhisek Ray, Ayush Raj, and Maheshkumar H Kolekar. Autoregressive adaptive hypergraph transformer for skeleton-based activity recognition. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9690–9699. IEEE, 2025.

[72] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 479–485, 2013.

[73] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[74] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[75] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[76] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[77] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13413–13422, 2021.

[78] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019.

[79] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[80] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[81] Yi-Fan Song, Zhang Zhang, and Liang Wang. Richly activated graph convolutional network for action recognition with incomplete skeletons. In *International Conference on Image Processing (ICIP)*, pages 1–5. IEEE, 2019.

[82] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017.

[83] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225, 2023.

[84] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[85] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

[86] Raviteja Vemulapalli and Rama Chellapa. Rolling rotations for recognizing human actions from 3d skeletal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4471–4479, 2016.

[87] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 499–508, 2017.

[88] Hongsong Wang and Liang Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018.

[89] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1290–1297. IEEE, 2012.

[90] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014.

[91] Lei Wang. Analysis and evaluation of Kinect-based action recognition algorithms. Master's thesis, School of the Computer Science and Software Engineering, The University of Western Australia, Nov 2017.

[92] Lei Wang. *Robust human action modelling*. PhD thesis, The Australian National University (Australia), 2023.

[93] Lei Wang, Du Q. Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. *TIP*, 29(1):15–28, 2019.

[94] Lei Wang, Du Q. Huynh, and Moussa Reda Mansour. Loss switching fusion with similarity search for video classification. In *IEEE ICIP*, pages 974–978, 2019.

[95] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4324–4333, 2021.

[96] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Proceedings of the Asian conference on computer vision*, pages 4176–4193, 2022.

[97] Lei Wang and Piotr Koniusz. Uncertainty-dtw for time series and sequences. In *European Conference on Computer Vision*, pages 176–195. Springer, 2022.

[98] Lei Wang and Piotr Koniusz. 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5620–5631, 2023.

[99] Lei Wang and Piotr Koniusz. Flow dynamics correction for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3795–3799. IEEE, 2024.

[100] Lei Wang and Piotr Koniusz. Feature hallucination for self-supervised action recognition. *International Journal of Computer Vision (IJCV)*, 2025.

[101] Lei Wang, Piotr Koniusz, and Du Q Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8698–8708, 2019.

[102] Lei Wang, Jun Liu, and Piotr Koniusz. 3d skeleton-based few-shot action recognition with jeanie is not so naïve. *arXiv preprint arXiv:2112.12668*, 2021.

[103] Lei Wang, Jun Liu, Liang Zheng, Tom Gedeon, and Piotr Koniusz. Meet jeanie: a similarity measure for 3d skeleton sequences via temporal-viewpoint alignment. *International Journal of Computer Vision*, 132(9):4091–4122, 2024.

[104] Lei Wang, Ke Sun, and Piotr Koniusz. High-order tensor pooling with attention for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3885–3889. IEEE, 2024.

[105] Lei Wang, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng. Taylor videos for action recognition. In *Forty-first International Conference on Machine Learning*.

[106] Peng Wang, Fanwei Zeng, and Yuntao Qian. A survey on deep learning-based spatio-temporal action detection. *International Journal of Wavelets, Multiresolution and Information Processing*, 22(04):2350066, 2024.

[107] Shengqin Wang, Yongji Zhang, Hong Qi, Minghao Zhao, and Yu Jiang. Dynamic spatial-temporal hypergraph convolutional network for skeleton-based action recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2147–2152. IEEE, 2023.

[108] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[109] Syed Talal Wasim, Muhammad Uzair Khattak, Muzammal Naseer, Salman Khan, Mubarak Shah, and Fahad Shahbaz Khan. Video-focalnets: Spatio-temporal focal modulation for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[110] Jiangning Wei, Lixiong Qin, Bo Yu, Tianjian Zou, Chuhan Yan, Dandan Xiao, Yang Yu, Lan Yang, Ke Li, and Jun Liu. Va-ar: Learning velocity-aware action representations with mixture of window attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8286–8294, 2025.

[111] Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, and Shihong Xia. Graph cnns with motif and variable temporal block for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8989–8996, 2019.

[112] Cong Wu, Xiao-Jun Wu, and Josef Kittler. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[113] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 20–27. IEEE, 2012.

[114] Chunyu Xie, Ce Li, Baochang Zhang, Chen Chen, Jungong Han, and Jianzhuang Liu. Memory attention networks for skeleton-based action recognition. In *International Joint Conference on Artificial Intelligence*, 2018.

[115] Jianyang Xie, Yanda Meng, Yitian Zhao, Nguyen Anh, Xiaoyun Yang, and Yalin Zheng. Dynamic semantic-based spatial-temporal graph convolution network for skeleton-based human action recognition. *IEEE Transactions on Image Processing*, 2024.

[116] Haojun Xu, Yan Gao, Zheng Hui, Jie Li, and Xinbo Gao. Language knowledge-assisted representation learning for skeleton-based action recognition. *IEEE Transactions on Multimedia*, pages 1–16, 2025.

[117] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2866–2874, 2022.

[118] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[119] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[120] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14063–14073, June 2022.

[121] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *International Conference on Multimedia (MM)*, pages 55–63, 2020.

[122] Lin Yuan, Zhen He, Qiang Wang, Leiyang Xu, and Xiang Ma. Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based tai chi action recognition. In *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2022.

[123] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012.

[124] Xiao Yun, Chenglong Xu, Kevin Riou, Kaiwen Dong, Yanjing Sun, Song Li, Kevin Subrin, and Patrick Le Callet. Behavioral recognition of skeletal data based on targeted dual fusion strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6917–6925, 2024.

[125] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017.

[126] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1112–1121, 2020.

[127] Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14333–14342, 2020.

[128] Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. Bayesian graph convolution lstm for skeleton based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6882–6892, 2019.

[129] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yanwen Fang, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.

[130] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefining topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[131] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024.

[132] Yisheng Zhu, Hui Shuai, Guangcan Liu, and Qingshan Liu. Multilevel spatial–temporal excited graph network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 32:496–508, 2023.