Computer Science Review 62 (2026) 101030

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Beyond benchmark accuracy: A generalization-centered analysis of deep learning for skeleton-based human activity recognition

Yi Xia [a], Sira Yongchareon [a,*], Raymond Lutui [a], Quan Z. Sheng [b]

[a] School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, 55 Wellesley Street East, Auckland, 1010, New Zealand
[b] School of Computing, Faculty of Science and Engineering, Macquarie University, Balaclava Road, Sydney, NSW 2109, Australia

ARTICLE INFO

Keywords:
Human activity recognition
Skeleton data
Deep learning
Generalization
Domain adaptation
Self-supervised learning

ABSTRACT

Deep learning models for skeleton-based Human Activity Recognition (HAR) routinely exceed 93% accuracy on controlled benchmarks such as NTU RGB+D, yet degrade by 30–60% when confronted with the noise, demographic diversity, and environmental variability of real-world deployment. Existing surveys catalog this progress by network architecture (RNNs, CNNs, GCNs, Transformers) but provide limited guidance on when or why one approach outperforms another under specific deployment constraints. This survey reframes the field through a *generalization-centered* lens. We identify and quantify six deployment challenges that drive performance degradation: sensor noise and pose estimation errors (46–60 percentage point gap between clean Kinect and RGB-estimated skeletons), open vocabulary and unseen actions (47–63 pp between supervised and zero-shot results), demographic shift (20–40 pp across elderly-inclusive benchmarks), cross-dataset transfer (10–33 pp), annotation scarcity for purely supervised training (10–15 pp), and cross-subject or cross-view variation on a single dataset (3–7 pp). For each challenge, we systematically evaluate how representation paradigms (sequential, spatial, graph, attention, and cross-modal), learning strategies (self-supervised, semi-supervised, few-shot, and domain adaptation), and integration techniques (fusion, distillation, and multi-task) perform, and we provide cross-method comparisons grounded in published results. Our analysis shows that no method effectively addresses more than three challenges, that research effort focuses on the smallest gaps while the largest remain understudied, and that self-supervised learning and CNN-based representations emerge as the most versatile paradigms across multiple deployment scenarios. We synthesize these findings into a method–challenge suitability matrix and a decision framework that links deployment constraints to method recommendations. These recommendations are derived from single-challenge evidence and require validation under simultaneous-challenge deployment conditions. A three-tier research roadmap identifies concrete directions for bridging the generalization gap across near-, mid-, and long-term horizons.

1. Introduction

The discrepancy between laboratory benchmarks and real-world performance remains a key unsolved challenge in skeleton-based Human Activity Recognition (HAR). While deep learning models achieve high accuracy (>93%) on curated datasets like NTU RGB+D [85], their performance degrades by 30–60% in field deployments [17,26]. This gap points to a core issue: incremental improvements on benchmark leaderboards often fail to translate into reliable, deployable solutions.

Skeleton-based HAR, which models human motion as a temporal sequence of joint coordinates [14], offers clear advantages for real-world applications. It is efficient and privacy-preserving by design compared to RGB video [9,91], and device-agnostic, enabling use in healthcare monitoring [32], smart environments [4], industrial safety [43], and human-computer interaction [18]. The field has produced rapid architectural evolution, from RNNs [24] and GCNs [122] to Transformers [80] and LLM-integrated systems [82], yet widespread adoption remains hindered by the recurring gap between benchmark accuracy and deployment reliability.

Previous surveys have provided valuable catalogs of this evolution, typically organized by network architecture [44,65,69,84,104] or by specific technical strands such as transformers [119] and self-supervised learning [130]. While these architectural taxonomies are useful for

* Corresponding author.
Email addresses: yi.xia@autuni.ac.nz (Y. Xia), sira.yongchareon@aut.ac.nz (S. Yongchareon), raymond.lutui@aut.ac.nz (R. Lutui), michael.sheng@mq.edu.au (Q.Z. Sheng).

https://doi.org/10.1016/j.cosrev.2026.101030
Received 1 March 2026; Received in revised form 7 June 2026; Accepted 27 June 2026
Available online 3 July 2026

navigation, they prioritize describing *what* methods exist over analyzing *when* and *why* particular approaches succeed or fail under specific deployment conditions. Emerging trends (LLM integration for zero-shot recognition [82,123], self-supervised methods that reduce annotation requirements by 90% [2]) are typically treated as isolated novelties rather than as responses to specific generalization challenges. As a result, practitioners seeking to deploy skeleton-based HAR in real-world scenarios receive limited guidance on method selection.

In this survey, we argue that the central question is not "which method achieves the highest benchmark accuracy?" but "which generalization challenges prevent deployment, and how effectively do different methods address each one?" We identify six generalization challenges that systematically degrade real-world performance, ordered by the severity of degradation they induce:

1. **Sensor noise and pose estimation errors**: The transition from clean depth-sensor skeletons to 2D-estimated poses from RGB video introduces errors that reduce accuracy from 93% on NTU RGB + D to 33–47% on Kinetics-Skeleton.
2. **Open vocabulary and unseen actions**: Supervised models cannot recognize actions absent from training data; reported zero-shot results on skeleton-only methods span 32–46% depending on the benchmark.
3. **Demographic shift**: Models trained on young-adult benchmarks degrade on elderly and mobility-impaired populations, with reported gaps spanning 20–40 pp across elderly-inclusive studies.
4. **Cross-dataset transfer**: Models trained on one dataset fail on others despite shared action vocabularies, indicating that they learn dataset-specific artifacts.
5. **Annotation scarcity**: Specialized domains (medical, sports, industrial) cannot afford the large-scale annotation that purely supervised methods require, although self-supervised pre-training closes this gap on standard benchmarks.
6. **Cross-subject and cross-view variation**: Standard NTU RGB + D protocols receive the largest share of research attention yet induce the smallest gaps.

For each challenge, we evaluate methods from all paradigms (representation, learning, and integration) within a unified analytical framework, enabling direct cross-method comparison that architecture-organized surveys cannot provide. This analysis produces several findings absent from prior surveys: a challenge severity ranking showing misallocated research efforts, a method–challenge suitability matrix exposing systematic coverage gaps, and a deployment decision framework mapping practical constraints to method recommendations.

The primary contributions of this work include:

- A *generalization-centered analytical framework* that identifies, quantifies, and ranks six deployment challenges, shifting evaluation from benchmark accuracy to deployment readiness.
- *Systematic cross-method comparisons* within each challenge, revealing which paradigms (SSL, CNN, GCN, DA) hold advantages under specific deployment conditions and where evaluation coverage is missing.
- A *method–challenge suitability matrix* and *deployment decision framework* that provide practical guidance for practitioners, mapping deployment scenarios to recommended method combinations.
- A *quantitative analysis of research effort allocation*, exposing an inverse relationship between challenge severity and evaluation coverage, with a three-tier roadmap targeting under-studied deployment barriers.

The remainder of this survey is organized as follows. Section 2 positions our work within the existing survey literature and identifies the gaps that our generalization-centered perspective addresses. Section 3 provides a concise taxonomy of the method landscape, covering representation paradigms, learning strategies, and integration approaches. Section 4 (the analytical core) systematically evaluates methods against six generalization challenges, producing cross-cutting comparisons and a practical decision framework. Section 5 analyzes datasets, evaluation protocols, and performance trends, connecting benchmark limitations to generalization failures. Section 6 outlines a research roadmap organized by unresolved challenges. Section 7 concludes with key findings and practical recommendations.

1.1. Scope and methodology of this review

We conducted a systematic literature search combining two major indexing databases (Scopus and Web of Science) for broad coverage of peer-reviewed publications, two publisher-specific databases (IEEE Xplore and ACM Digital Library) for targeted retrieval of top venues in computer vision and pattern recognition, the DBLP computer science bibliography for author- and venue-level verification, Google Scholar for citation analysis and grey literature discovery, and arXiv for recent preprints not yet indexed elsewhere. The search used combinations of keywords including "skeleton-based action recognition," "human activity recognition," "pose estimation + activity," "graph convolutional network + skeleton," "transformer + skeleton action," and "self-supervised + skeleton." We restricted the primary search window to 2016–2026, corresponding to the deep learning era in skeleton-based HAR initiated by ST-LSTM [62] and ST-GCN [122], while including foundational works from earlier years where necessary for context.

Inclusion criteria required that papers (1) address skeleton-based or pose-based human activity recognition using deep learning, (2) present empirical evaluation on at least one established benchmark, and (3) be published in peer-reviewed venues or deposited on arXiv with subsequent citation uptake. We excluded papers focused exclusively on pose estimation without recognition, RGB-only activity recognition without skeletal components, and non-deep-learning approaches (e.g., handcrafted feature methods). For rapidly evolving topics such as LLM integration and self-supervised learning, we included preprints from 2025 to 2026 that had gained community attention. Duplicate records across databases were removed during the consolidation stage.

This process yielded approximately 200 references spanning five representation paradigms (sequential, spatial, graph-based, attention-based, and cross-modal), four learning strategies (data augmentation, self-supervised learning, limited-label learning, and cross-domain transfer), and three hybrid integration approaches (multi-component fusion, multi-task learning, and knowledge distillation). The distribution reflects the field's trajectory: GCN-based methods constitute the largest group, followed by transformer and self-supervised approaches, with LLM-integrated methods representing the most recent and fastest-growing category.

2. Related work

Skeleton-based HAR has attracted numerous surveys, each contributing to a partial understanding of the field. Table 1 compares representative prior surveys along seven dimensions, including a new assessment of which generalization challenges each survey addresses. This comparison shows that while architectural coverage has expanded considerably, systematic analysis of deployment-relevant generalization challenges remains absent.

Architecture-centric reviews constitute the most common category and provide essential technical catalogs. Wang et al. [104] compare Kinect-era algorithms; Kong and Fu [44] extend the scope to action prediction; Ren et al. [84] offer a chronological catalog spanning RNNs, CNNs, GCNs, and Transformers. More recently, Liu et al. [65] adopt a task-oriented decomposition covering preprocessing, feature extraction, and spatiotemporal modeling, including Mamba and LLM-based architectures. Liu et al. [69] identify three challenges (labeled data scarcity, few-shot recognition, and missing skeleton information) in a systematic review. These works document *what* models exist and how they perform on standard benchmarks but do not systematically analyze *when*

Table 1
Comparison of existing surveys on skeleton-based human activity recognition. Challenge coverage columns indicate whether each survey systematically analyzes the corresponding generalization challenge: **N** = Noise robustness, **S** = Cross-subject/demographic, **V** = Viewpoint/environment, **L** = Label scarcity, **O** = Open vocabulary, **E** = Computational efficiency.

Reference	Year	Venue	Primary Focus	Repr.	Learn.	Gen.	Chall.	Unique Contribution
Wang et al. [104]	2019	TIP	Kinect-based methods	✓			–	Kinect-era algorithm comparison
Kong & Fu [44]	2022	IJCV	Action recognition & prediction	✓	Partial		–	Prediction task integration
Yue et al. [128]	2022	Neurocomputing	RGB vs. skeleton	Partial			–	Cross-modality comparison
Sun et al. [94]	2022	TPAMI	Multi-modal HAR	Partial			–	Modality taxonomy
Xin et al. [119]	2023	Neurocomputing	Transformers for skeleton	Partial			E	Transformer architecture analysis
Ren et al. [84]	2024	Cyborg Bionic Sys.	3D skeleton learning	✓	Partial		–	Chronological method catalog
Dalal & Mittal [20]	2025	Comput. Sci. Rev.	DL for elderly & human AR	Partial	Partial		S	PRISMA-based; elderly focus
Liu et al. [69]	2025	IEEE TNNLS	Skeleton AR challenges	✓	Partial		N, L	Three-challenge decomposition
Zhang et al. [130]	2026	IJCV	Self-supervised skeleton	Partial	✓	✓	N, S	SSL benchmark; cross-dataset eval
This survey	**2026**		**Generalization-centered**	✓	✓	✓	**All 6**	**Challenge severity ranking; method–challenge matrix; deployment framework**

Repr.: Coverage of multiple representation paradigms. *Learn.*: Analysis of learning strategies (SSL, few-shot, domain adaptation). *Gen.*: Cross-dataset generalization assessment. *Chall.*: Specific generalization challenges addressed (N = Noise, S = Subject/demographic, V = Viewpoint, L = Label scarcity, O = Open vocabulary, E = Efficiency).

or *why* particular paradigms succeed or fail under specific deployment conditions.

Technology-focused and application-oriented surveys examine specific technical strands or domains. Xin et al. [119] analyze Transformer adaptations for skeletons; Yue et al. [128] and Sun et al. [94] compare skeleton with RGB and depth modalities. Zhang et al. [130] present a benchmark for self-supervised skeleton representation learning in IJCV, uniquely including cross-dataset generalization assessment. This is the closest prior work to our generalization-centered perspective, though limited to SSL methods. Dalal and Mittal [20] provide a PRISMA-based review in this journal on elderly activity recognition, addressing demographic generalization within a single application context (Fig. 2).

As shown in Table 1, three systematic gaps emerge despite growing coverage:

First, no prior survey jointly and systematically analyzes all six generalization challenges. Architecture-centric reviews [44,65,69,84,104] cover representation paradigms but give limited attention to domain adaptation, demographic shifts, or computational efficiency. The closest works address individual challenges in isolation: Liu et al. [69] discuss label scarcity and missing joints; Zhang et al. [130] evaluate cross-dataset transfer for SSL; Dalal and Mittal [20] address elderly populations. No survey provides a unified analysis comparing how *different* method families handle the *same* challenge.

Second, no prior survey quantifies challenge severity or ranks generalization gaps. The 46–60 percentage point gap caused by sensor noise and the 10–33 percentage point gap in cross-dataset transfer have been documented in individual papers [26,36] but have not been systematically compared to show where the field's efforts are most versus least productively allocated.

Third, no prior survey provides practical deployment guidance. Practitioners face specific combinations of challenges (e.g., noisy data + limited labels + real-time constraints for hospital monitoring) and need method recommendations tailored to these combinations. Architecture-organized surveys cannot provide this guidance: methods are described within paradigm silos rather than compared on common deployment challenges.

This survey addresses all three gaps. Our generalization-centered analysis evaluates methods from all paradigms within each challenge, producing cross-cutting comparisons that expose method suitability patterns invisible to architecture-organized reviews. The resulting method–challenge suitability matrix (Table 4) and deployment decision framework (Table 5) translate analytical findings into practical guidance.

3. Method landscape: a concise taxonomy

Before analyzing generalization challenges, we provide a concise taxonomy of the methodological toolkit available for skeleton-based HAR. This section serves as a reference map: it introduces the key method families, their core mechanisms, and representative approaches. The challenge-driven analysis of these methods, evaluating when and why each succeeds or fails, is deferred to Section 4.

Fig. 1 illustrates the taxonomy. Methods are organized into three categories: representation paradigms (how skeletal data is encoded), learning strategies (how models are trained under constraints), and integration approaches (how complementary components are combined). Table 2 summarizes trade-offs across representation paradigms.

3.1. Representation paradigms

Sequential Representations (RNNs). Recurrent neural networks encode skeleton sequences via temporal hidden states, a design well-suited to causal (online) processing. Representative variants include hierarchical RNNs [24], ST-LSTM with trust gates [62], and deeper IndRNN architectures [134], with attention refinements at the joint, frame, and neuron levels [63,93,132] and graph-augmented recurrent designs such as AGC-LSTM [89]. Their peak accuracy on the NTU RGB + D cross-view dataset reaches 88.4% [132], which remains 7–12 percentage points below graph- and attention-based alternatives due to limitations in sequential computation and gradient propagation.

Spatial grid representations (CNNs). Convolutional networks turn skeletons into grid-structured inputs for parallel hierarchical feature extraction, originally as 2D trajectory images [34,107] and later as multi-stream and 3D variants [42,52,59]. PoseC3D [26] represents skeletons as 3D heatmap volumes and achieves 93.7% on NTU and 47.7% on Kinetics-Skeleton, illustrating how the grid format provides implicit smoothing over pose estimation noise. Section 4.2 analyzes this advantage compared to graph-based alternatives, and Section 4.7 discusses

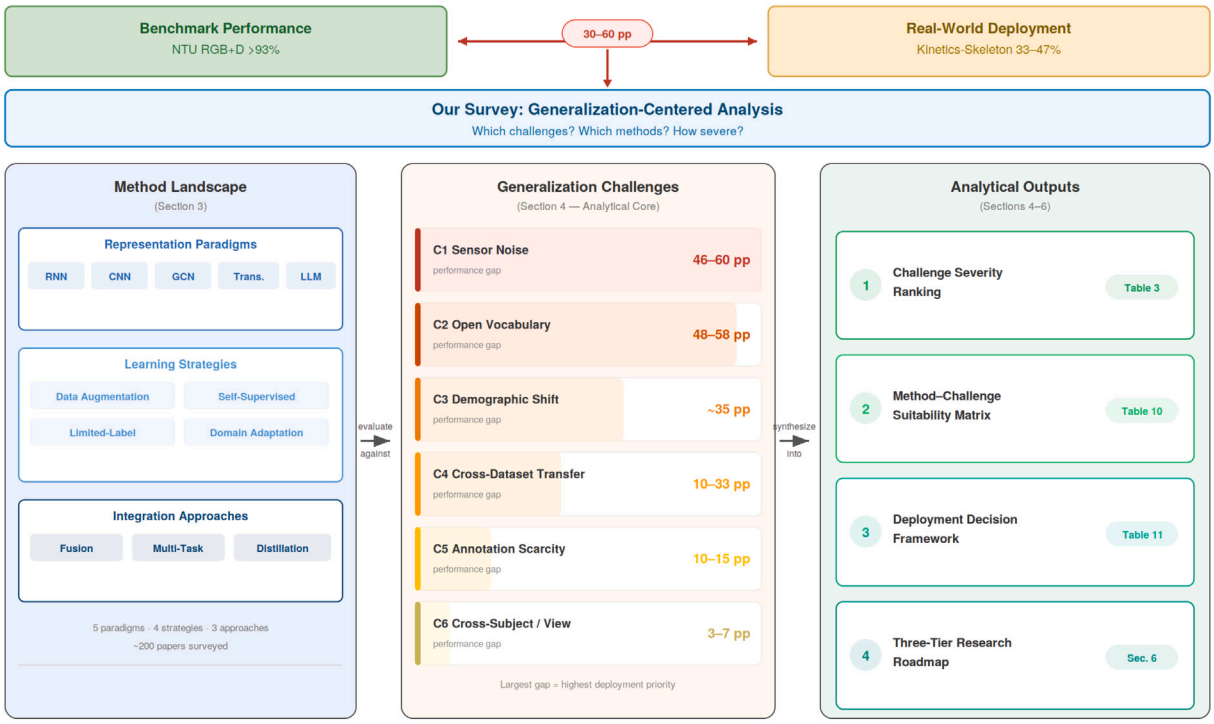

Fig. 1. Overview of the analytical framework. Section 3 provides a concise taxonomy of the methods landscape (representation paradigms, learning strategies, integration approaches). Section 4 evaluates these methods through six generalization challenges, producing cross-cutting analyses and a practical decision framework.

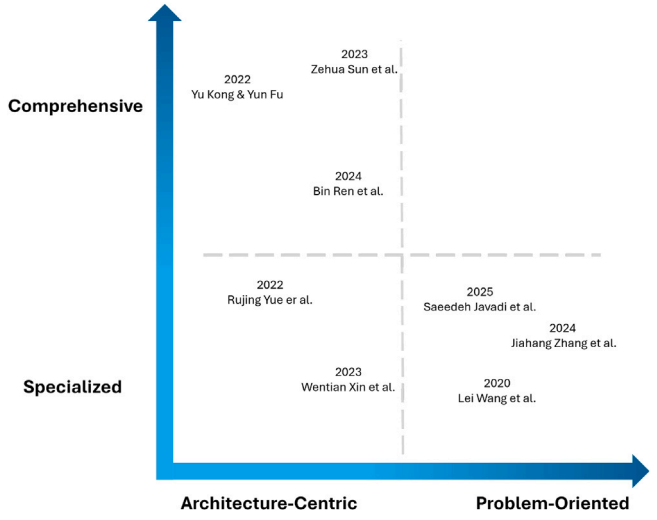

Fig. 2. Positioning of existing surveys on skeleton-based HAR along two dimensions: organizational focus (architecture-centric vs. problem-oriented) and technical scope (specialized vs. comprehensive). This survey occupies the problem-oriented, comprehensive quadrant by analyzing all method families through the lens of generalization challenges.

pruning and depthwise separable designs that enable real-time inference [76].

Graph-structured representations (GCNs). Graph convolutional networks encode skeletal topology explicitly by treating joints as nodes and bones as edges. ST-GCN [122] introduced the spatial-temporal graph formulation, and subsequent work has improved accuracy through adaptive topology learning [16,87], semantic-guided construction [12,30], hybrid attention or recurrent coupling [22,115], anatomically-informed designs [41,66,118], and efficiency refinements [81,137]. On clean NTU RGB + D they reach up to 97.5% cross-view and retain strong sample efficiency, making them the dominant paradigm on controlled benchmarks; their behavior under noise is analyzed in Section 4.2.

Attention-based representations (Transformers). Transformers model joint-frame dependencies through self-attention without predefined connectivity. Starting from dual spatial-temporal designs [80], recent work tackles the $O(N^2T^2)$ cost through anatomical partitioning [23], frequency-domain mixing [114], local-global hybrids [67], and graph-transformer crosses [135]. Peak accuracy reaches 98.3% on NW-UCLA [23], but attention-based models need more than 100K training samples to outperform GCNs and still exceed 200 ms inference latency; Section 4.7 analyzes this constraint.

Cross-modal semantic representations (LLMs). LLM-based methods transfer external semantic knowledge to skeleton HAR through cross-modal alignment. Auxiliary designs use LLM-generated descriptions to guide training via contrastive or attention losses [117,123], while direct-integration designs encode skeletons into LLM token spaces: LLM-AR [82] achieves 95.0% on NTU after LoRA fine-tuning, and SKI [92] projects skeletons into CLIP's embedding space. Language-assisted distillation such as LA-GCN [120] retains most of the semantic benefits without the LLM inference cost. This paradigm uniquely supports open vocabulary recognition; Section 4.6 discusses its 200–500 ms latency and cultural-bias trade-offs.

3.2. Learning strategies

Data augmentation and synthesis. Generative augmentation expands limited training sets while preserving biomechanical plausibility.

Table 2
Comparative summary of representation paradigms for skeleton-based HAR. Peak accuracy is on NTU RGB+D 60 cross-subject unless noted; latency, FLOPs, and other quantitative claims follow the cited references.

Paradigm	Core mechanism	Reported strengths	Reported weaknesses	Indicative settings	Peak Acc.
Sequential (RNN)	Temporal hidden states with causal sequential updates [62,132,134]	Native streaming inference; small per-step memory [62,132]	7–12 pp below GCNs and transformers; vanishing gradients on long sequences [132]	Online/causal pipelines, low-power streaming	81.8% [134]
Spatial grid (CNN)	2D/3D convolutions over heatmap or trajectory representations [26,34,76]	Implicit smoothing of pose-estimation noise (47.7% on Kinetics) [26]; 4.10 ms inference on edge ResNet [76]	Loses explicit anatomical structure; limited modeling of long-range joint coupling [26]	Pose-estimated input; latency-bounded edge devices	93.7% [26]
Graph (GCN)	Message passing on a fixed or adaptive skeletal graph [16,66,122]	Strong sample efficiency from anatomical prior (90% accuracy at ~50K samples) [16]	60–70% accelerator utilization on sparse graph ops [90]; error propagation under noise [97,122]	Clean Kinect data, small-to-medium training sets	93.7% [12]
Attention (Trans.)	Self-attention over joint-frame tokens [23,80,114]	Long-range dependency modeling; best NW-UCLA accuracy 98.3% [23]	$O(N^2T^2)$ cost; >200 ms latency on 300-frame inputs; requires ≥100K training samples [16,23]	Large-scale offline analysis	93.6% [114]†
Cross-modal (LLM)	Skeleton-text alignment via prompts, embeddings, or token quantization [82,92,117]	Open-vocabulary recognition; highest reported NTU60 accuracy 95.0% [82]	200–500 ms latency; reliance on curated text descriptions; cultural bias of LLM corpora [82,92]	Zero-shot recognition; semantic disambiguation	95.0% [82]

† Transformer peak on NTU 60 cross-subject is 93.6 % [114]; the family-best result, 98.3 %, is reported on NW-UCLA [23]. Differences within ±0.3–0.5 pp reflect run-to-run variance [85].

Conditional generators [21,86,109], view-normalizing GANs [78,83], and occlusion-focused reconstructors [100] address complementary data gaps. Quantitative evaluations appear in Sections 4.5 and 4.2.

Self-supervised representation learning. Self-supervised learning extracts transferable features without labels through pretext objectives. Three families dominate: contrastive methods [2,68], predictive or JEPA-style methods [1], and disentanglement approaches [55,112]. Multi-modal SSL [35] combines skeletal data with complementary modalities. Their behavior under noise, demographic shifts, and annotation scarcity is analyzed in Sections 4.2–4.5.

Learning from limited labels. Three families address annotation scarcity. *Semi-supervised learning* combines limited labels with unlabeled data [37,71,99,136], *active learning* selects informative samples for annotation [47,133], and *few-shot learning* enables rapid adaptation with a handful of examples per class [10,48,138]. Sections 4.5 and 4.6 report their quantitative performance and failure modes.

Cross-domain transfer. Transfer learning and domain adaptation address distribution shifts between training and deployment. Transfer methods adapt pre-trained models through fine-tuning or prompt tuning [25,70,124], while domain adaptation methods align source and target distributions through cross-attention, multimodal alignment, graph heterogeneity handling, or contrastive-MMD combinations [36,75,97,121]. Sections 4.3 and 4.4 analyze these methods under specific generalization settings.

3.3. Integration approaches

Multi-component integration. Multi-stream fusion combines diverse skeleton representations or modalities [60,88], while model ensembles aggregate complementary architectures (GCN and Transformer branches, front-rear fusion, decoupled general and individual graph components) to balance accuracy and efficiency [64,129,139].

Multi-task learning. Joint optimization of related objectives provides implicit regularization. Generic combinations integrate motion prediction, jigsaw recognition, and contrastive learning [54], while domain-specific designs pair action recognition with depth estimation for surveillance, quality evaluation for fitness, or static-dynamic correlation decoupling [6,106,141].

Knowledge distillation. Distillation transfers knowledge from complex teachers to lightweight students. Approaches include temporal-response distillation for efficiency [140], cross-modal distillation that keeps skeletons as the only inference input [96], multi-teacher designs for occlusion robustness [126], part-level distillation for noisy inputs [57], and format-adaptive variants [108]. Sections 4.7 and 4.2 examine their effects on the corresponding deployment constraints.

4. Generalization challenges: a systematic analysis

The method landscape presented in Section 3 provides a toolkit of representation paradigms, learning strategies, and integration approaches. This section, the analytical core of our survey, evaluates these approaches through the lens of generalization: the recurring gap between benchmark performance and real-world deployment. Rather than asking "what does each method do?" we ask "what challenges prevent deployment, and which methods address them?"

We identify six generalization challenges that systematically degrade field performance (Table 3). These challenges are not independent: real deployments typically face several of them at once (Section 4.8). Our analysis reveals a gap between current research and deployment needs: single-challenge benchmarks do not reflect the combinations a real system must handle, as Table 4 quantifies in Section 4.8.

4.1. Taxonomy and severity of generalization challenges

The six challenges share a single organizing principle: each is a measurable source of the accuracy gap between controlled benchmarks and real-world deployment. This grouping is deliberately outcome-oriented rather than a taxonomy of causes, so the challenges arise from different mechanisms. Some are distribution shifts encountered at deployment (sensor noise, demographic shift, cross-dataset transfer), one is a shift in the action vocabulary itself (open vocabulary and unseen actions), one is a learning constraint (annotation scarcity), and the last is a property of the evaluation protocol (cross-subject and cross-view splits). We adopt this pragmatic organization because practitioners face these factors together as deployment barriers, regardless of their underlying cause. A direct consequence is that the categories are not mutually exclusive. Cross-subject evaluation is both a protocol and a limited proxy for demographic shift; cross-view variation overlaps with the broader cross-dataset and environment shift; and sensor noise compounds annotation scarcity when noisy data also lack reliable labels. Accordingly,

Table 3
Generalization challenge severity ranking for skeleton-based HAR, ordered by the magnitude of performance degradation each challenge induces. Gap ranges are based on published experimental results, with cross-protocol caveats summarized in the table notes.

Challenge	Representative evidence	Gap	References
Sensor noise & pose estimation errors	NTU RGB + D 93% (Kinect) vs. Kinetics-Skeleton 33–47% (RGB-estimated); controlled Gaussian noise on clean NTU skeletons adds 15–30 pp alone	46–60 pp[†]	[26,58,79,122]
Open vocabulary & unseen actions	Supervised 93–95% vs. zero-shot 32–46% (PURLS on Kinetics-skeleton 200 to SA-DVAE on NTU 120)	47–63 pp[‡]	[10,48,138]
Demographic shift (age, body type, ability)	Cross-age gaps reported across elderly-inclusive benchmarks span 20–40 pp; no standard protocol pairs young and elderly cohorts with matched action vocabularies	20–40 pp[§]	[20,32,39]
Cross-dataset transfer	NTU → PKU-MMD: 33.5 pp drop (51.6% vs. 85.1% target-trained); NTU → NW-UCLA: 10–28 pp	10–33 pp	[36,97,121]
Annotation scarcity (supervised)	Purely supervised training on 10% of NTU labels loses 10–15 pp relative to 100% labels; self-supervised pre-training closes this gap (Section 4.5)	10–15 pp	[2,71,112]
Cross-subject/cross-view	Within-dataset protocol variations on NTU RGB + D	3–7 pp	[61,85]

[†] Combines pose estimation noise (15–30 pp in controlled NTU experiments [79]) and dataset-difficulty increase (60 vs. 400 classes between NTU and Kinetics); the two effects are inseparable in deployment.
[‡] Zero-shot ranges combine PURLS [138] on Kinetics-skeleton 200, STAR [10] and SA-DVAE [48] on NTU 60/120; protocols differ across papers.
[§] No single paper reports a controlled young-vs-elderly comparison with matched action vocabularies; the range summarizes elderly-HAR results synthesized in [20] together with ETRI-Activity3D [39] and earlier elderly-specific studies [32].

the severity ranking (Table 3) reports gaps at a finer granularity, separating demographic shift and cross-dataset transfer, whereas the suitability matrix (Table 4) uses the coarser C1–C6 grouping and adds computational efficiency (C6), a latency constraint rather than an accuracy gap. We therefore classify each method by the dominant challenge it targets, and we flag these overlaps where they affect interpretation. Alternative groupings, for example by type of distribution shift, are possible and would offer a complementary view.

Table 3 quantifies six generalization challenge types by the performance degradation they induce, ordered by severity. This ranking highlights an apparent mismatch between challenge severity and research attention: the majority of recent publications target cross-subject and cross-view settings where gaps are smallest (3–7 percentage points), while sensor noise and cross-dataset transfer (inducing 30–60 point gaps) receive comparatively less attention.

The severity ranking carries practical implications. Sensor noise alone accounts for the largest performance gap in the field, yet most architectural innovations are evaluated exclusively on clean Kinect-captured benchmarks (NTU RGB + D) where this challenge is absent. Open vocabulary recognition and demographic shift produce comparably large gaps but are addressed by fewer than 10% of the methods surveyed. In contrast, cross-subject and cross-view generalization (the standard evaluation protocols for NTU RGB + D) induce the smallest gaps and have received the most attention, approaching saturation with over 10 methods achieving 93–95% accuracy.

These severity rankings should be read as indicative rather than exact. As the notes to Table 3 detail, each range aggregates results obtained on different datasets, skeleton sources, and evaluation protocols, and few papers report controlled single-variable comparisons. Comparing the ranges therefore rests on three assumptions: that the reported accuracies refer to comparable label spaces, that the named shift is the dominant source of degradation rather than a confound, and that protocol differences across papers are smaller than the gaps used to rank them. The separation between the largest and smallest gaps holds under these assumptions, but the precise widths, and the ordering of challenges with overlapping ranges such as demographic shift and cross-dataset transfer, warrant corresponding caution.

The following subsections analyze each challenge in depth, evaluating how different method families perform and identifying which approaches offer genuine advantages versus incremental improvements.

4.2. Robustness to sensor noise and pose estimation errors

4.2.1. Problem definition and evidence

The transition from depth-sensor-captured skeletons (Microsoft Kinect) to 2D-pose-estimated skeletons (from RGB video) introduces three noise types: joint localization errors from occlusion and depth ambiguity, temporal jitter from frame-by-frame estimation inconsistencies, and missing joints from failed detection. This transition produces the largest performance gap in the field: methods achieving 93% on NTU RGB + D (Kinect-captured) degrade to 33–47% on Kinetics-Skeleton (pose-estimated from YouTube videos) [26,58,122].

The gap is not only a matter of dataset difficulty. Kinetics contains 400 action classes versus NTU's 60, and its unconstrained capture introduces background and lighting variation. Controlled experiments on NTU isolate the noise component: when clean Kinect skeletons are corrupted with Gaussian perturbations, methods degrade by 15–30 percentage points [79]. The remaining portion of the 46–60 pp gap arises from the class-count difference and the broader scene variation between the two benchmarks, and the two effects are inseparable in deployment because real systems face both jointly.

4.2.2. Method analysis across paradigms

CNN heatmap representations smooth noise implicitly. PoseC3D [26] represents the strongest approach to noise robustness by converting skeleton sequences into 3D heatmap volumes rather than processing raw joint coordinates. Each joint is represented as a Gaussian blob in a spatial grid, distributing positional information across neighboring voxels. This implicit spatial smoothing explains PoseC3D's 47.7% on Kinetics-Skeleton, a 10–14 percentage point advantage over all GCN variants. The key insight is that heatmap representations trade precision for robustness: minor localization errors shift the Gaussian center without eliminating the joint's signal, whereas coordinate-based methods treat each error as a discrete displacement. Efficiency-optimized CNN variants [76] achieve 4.10ms inference and 98.35% accuracy on clean data through depthwise separable convolutions and pruning, demonstrating that noise robustness need not sacrifice computational efficiency.

GCN message passing amplifies noise through neighborhood aggregation. Graph convolutions aggregate features from neighboring joints via message passing: $\mathbf{h}'_v = \sigma(\sum_{u \in \mathcal{N}(v)} \mathbf{W}\mathbf{h}_u)$. When a joint estimate is corrupted, its error propagates to all neighbors in the first layer and to second-order

neighbors in the next, creating an expanding contamination zone. This explains GCN methods' poor Kinetics performance: ST-GCN reaches only 33.4% [122] and SCA-GCN 37.3% [58], well below PoseC3D's 47.7%.

Adaptive topology learning [16,87], which replaces fixed skeletal adjacency with learned connections, exacerbates the problem. Analysis of learned adjacency matrices shows that data-driven topologies encode noise patterns specific to particular pose estimators rather than generalizable motion structures [97]. Fixed anatomical graphs transfer better across noise conditions than adaptive ones, suggesting that structural priors provide regularization against noise-induced spurious correlations. Semantic-guided approaches such as MAR-GCN [30], which construct hyperedges through meta-action semantics, and dependency refinement methods using Gaussian correlation functions [12] partially mitigate this by constraining the graph structure, but neither has been evaluated on noisy benchmarks.

Self-supervised pre-training learns noise-invariant representations. SSL methods show unexpected effectiveness against noise. MaskCLR [2] achieves 44.7% on Kinetics-Skeleton under linear evaluation, approaching PoseC3D's supervised 47.7% despite using no labeled data during pre-training. Its Attention-Guided Probabilistic Masking forces the model to reconstruct occluded joints from visible context, effectively training on a denoising objective. S-JEPA [1] shows even more direct evidence: when switching pose estimators between pre-training and evaluation (a controlled noise shift), generalization improves by 26.6% compared to supervised baselines, confirming that SSL captures motion semantics rather than estimator-specific artifacts.

IGM [55] achieves 63.0% accuracy under substantial synthetic noise by generating diverse augmented representations through diffusion processes with idempotency constraints, while SCD-Net [112] disentangles spatial and temporal features to prevent noise in one dimension from corrupting the other.

Occlusion completion and part-level distillation target specific noise types. GAN-based approaches address occlusion directly: Vernikos and Spyrou [100] use convolutional-recurrent generators to reconstruct missing joints, outperforming regression-based interpolation. Part-level knowledge distillation [57] transfers semantic body-part grouping from clean to noisy representations, improving robustness by learning to infer missing patterns from partial observations. MKE-GCN [126] uses multiple specialized teachers (joints, bones, motion) to supervise a unified student, improving robustness to viewpoint variations and occlusions.

4.2.3. Cross-method comparison and key insight

The noise robustness analysis exposes a core tension between structural priors and noise tolerance. GCNs' explicit encoding of skeletal topology provides a strong inductive bias on clean data (93–97% on NTU) but becomes a liability under noise because message passing treats all inputs (clean and corrupted) equally. CNNs' implicit spatial processing sacrifices structural specificity but gains noise resilience through spatial smoothing. SSL bridges this gap by learning noise-invariant features regardless of the underlying architecture.

Taken together, these results have a clear practical implication. For any deployment that relies on pose estimation from RGB video, which is the dominant real-world scenario, practitioners should prefer CNN-based or SSL-augmented representations over pure GCN approaches, despite the GCN advantage on clean data. This contradicts the current research trend in which GCN-based methods dominate publication venues based on NTU RGB + D results.

4.2.4. Open challenges

No method bridges the full 46-point gap between clean and noisy conditions. The 47.7% ceiling on Kinetics persists across all paradigms, suggesting an underlying information-theoretic limit: skeleton-only representations cannot disambiguate kinematically similar actions (e.g., "drinking" vs. "phone calling") that are distinguishable primarily by object context, absent in skeletal data. Whether joint confidence scores from pose estimators can be integrated into graph convolutions to weight message passing by reliability remains unexplored. Physics-informed approaches that reject biomechanically implausible joint configurations could provide principled noise filtering but have not yet been applied to this challenge.

4.3. Cross-subject and demographic generalization

4.3.1. Problem definition and evidence

Human motion varies widely across individuals due to differences in body proportions, movement habits, physical ability, age, and cultural background. Cross-subject evaluation protocols (e.g., NTU RGB + D: 20/20 subject split) test whether models generalize to unseen individuals, producing 3–7 percentage point gaps relative to cross-view settings. However, these protocols underestimate real-world demographic shifts because benchmark subjects are predominantly healthy young adults aged 18–35 recruited from university populations [61,85].

The severity of demographic shift becomes apparent in underrepresented populations. Reported cross-age gaps span 20–40 percentage points across elderly-inclusive benchmarks [20,32,39], reflecting differences in movement velocity, range of motion, and action execution. This range is considerably broader than the within-dataset cross-subject gap (3–7 pp) because age-related differences affect biomechanics rather than stylistic variation alone. ETRI-Activity3D [39] partially addresses this with 100 subjects but retains controlled laboratory conditions, and no published benchmark pairs young and elderly cohorts with matched action vocabularies. The core problem is that cross-subject protocols within homogeneous datasets test style variation within a narrow demographic, not true demographic generalization.

4.3.2. Method analysis across paradigms

Subject normalization reduces superficial variation. View-adaptive RNNs [131] normalize inputs to canonical orientations via learned transformations, improving cross-subject performance by eliminating viewpoint-dependent subject appearances. VN-GANs [78,83] standardize sequences across perspectives using adversarial training, minimizing individual variability while preserving action semantics. These approaches address the shallow end of subject variation (body orientation and global position) but cannot normalize deeper biomechanical differences such as joint range of motion or movement velocity profiles characteristic of different age groups or physical conditions.

Domain adaptation bridges demographic distributions. Adversarial domain adaptation explicitly aligns feature distributions between populations. Liao et al. [53] combine adaptive graph convolutions with gradient reversal layers to bridge source and target subject domains. STT-DA [121] uses cross-attention to align domains via parallel flows: self-attention for domain-specific features, cross-attention for inter-domain alignment, and bidirectional normalization to harmonize joint structures, achieving 82.5% on NTU-to-NW-UCLA transfer. However, these methods assume access to unlabeled target-domain data during training, which may not be available for underrepresented populations (e.g., patients with movement disorders).

Self-supervised learning captures subject-invariant motion primitives. SSL methods show particular promise for cross-subject generalization. S-JEPA [1] improves generalization by 26.6% when switching pose estimators, suggesting that its learned representations capture motion semantics rather than subject-specific or sensor-specific artifacts. Contrastive methods like MaskCLR [2] and SG-CLR [68] learn embeddings invariant to augmentation-induced variations (temporal jitter, joint noise) that partially overlap with inter-subject variation. SkeleMoCLR [71] combines momentum contrastive learning with teacher–student knowledge transfer to learn subject-invariant features for small-cohort datasets.

GCN structural priors provide sample-efficient generalization. GCNs' explicit anatomical encoding offers a structural advantage for cross-subject generalization: the fixed skeletal topology constrains the model to learn motion patterns along anatomically valid paths rather than subject-specific pose configurations. GCNs require approximately 50K training samples to reach 90% accuracy compared to 100K + for transformers [16], an advantage attributed to the graph prior reducing the effective hypothesis space. InfoGCN [16] improves on this by using information bottleneck principles to retain action-discriminative features while discarding subject-specific noise.

4.3.3. Cross-method comparison and key insight

Subject generalization benefits from complementary strategies operating at different levels. Normalization (VN-GAN) addresses superficial variation in body orientation; structural priors (GCNs) constrain learning to anatomically grounded patterns; SSL captures invariant motion primitives; and domain adaptation explicitly aligns population-level distributions. The most promising pipeline combines SSL pre-training (invariant features) with domain adaptation (population-level alignment), but fewer than five skeleton-HAR papers have explored this combination.

The key gap is *extreme* demographic shift: from healthy adults to patients with movement disorders (Parkinson's, stroke rehabilitation), where action execution patterns differ qualitatively, not just quantitatively. Current methods assume the same underlying motion vocabulary with stylistic variations; however, patient populations may exhibit qualitatively different movement primitives.

4.3.4. Open challenges

No benchmark provides controlled demographic diversity with ground-truth annotations. AddBiomechanics [111] offers biomechanically rich data from 273 subjects but covers only 9 activity classes. Constructing a cross-demographic benchmark with matched action vocabularies across age groups, body types, and ability levels is a prerequisite for systematic evaluation. Federated learning approaches that train across demographically diverse sites without centralizing sensitive health data represent a promising but largely unexplored direction, with current methods incurring 8–12% accuracy loss under differential privacy constraints [3].

4.4. Viewpoint and environmental adaptation

4.4.1. Problem definition and evidence

Deployment environments differ from training conditions in camera placement, background, lighting, and spatial configuration. Cross-view protocols on NTU RGB + D (3 camera angles) show only 1–3 percentage point gaps within the dataset, but cross-dataset transfer, which combines viewpoint shift with sensor change and population differences, produces 10–33 point degradation: NTU to NW-UCLA yields 57–86% depending on the method (versus 96–98% target-trained), while NTU to PKU-MMD yields only 51.6% (versus 85.1% target-trained) [36,97,121].

These gaps indicate that models learn environment-specific cues rather than view-invariant motion patterns. Learned graph topologies in adaptive GCNs encode camera-specific noise distributions from Kinect V2 depth estimation [97], while CNN-based methods exploit background regularities present in controlled laboratory settings. The gap persists even for shared action vocabularies, confirming that the problem is representational rather than semantic.

4.4.2. Method analysis across paradigms

Motion retargeting learns view-invariant representations. ViA [124] addresses viewpoint dependence directly through motion retargeting: learning to map sequences observed from one viewpoint onto a canonical skeleton representation using unlabeled data from multiple views. This self-supervised approach decouples action semantics from viewpoint-dependent appearance, achieving 78.1% on NTU cross-subject without explicit view labels. The approach shows that view-invariance can be learned as a pretext task rather than engineered manually.

Cross-attention domain adaptation aligns heterogeneous environments. STT-DA [121] applies cross-attention between source and target domains through parallel processing flows, achieving 82.5% on NTU-to-NW-UCLA transfer, closing the 28-point gap between no adaptation (57%) and target-trained performance (96%) by roughly half. MM-SADA [75] extends this to multimodal settings, integrating RGB, optical flow, and skeleton data with adversarial alignment, outperforming adversarial-only baselines by 3% on EPIC-Kitchens. The multimodal approach leverages a key insight: modality-specific domain shifts partially cancel when fused, as viewpoint changes affect RGB appearance more than skeletal geometry.

Graph heterogeneity alignment addresses structural differences. Cross-dataset transfer involves not only distributional shift but also structural differences: NTU uses 25-joint Kinect skeletons while COCO-based estimates yield 17-joint configurations. Tian et al. [97] address this with dual-space adversarial frameworks that decouple action semantics from graph-specific features, achieving a 17.2% improvement across graph-heterogeneous datasets. CStrCRL-UDA [36] combines maximum mean discrepancy with class-wise contrastive learning, reaching 71.3% on NTU60-to-PKUMMD transfer. ACFL [108] enables GCNs to adapt to diverse skeletal formats from single-format training data.

Transfer learning accelerates adaptation to new environments. Pre-trained models provide strong initialization for new environments. SkeleTR [25], a Transformer-GCN hybrid pre-trained on Posetics, transfers effectively to smaller benchmarks, reaching 94.8% on NTU. Skeleton-Prompt [70] introduces visual prompt tuning that freezes pre-trained weights and adapts via a skeleton prompt generator with cross-attention, achieving 93.6% on NTU 60 and 90.2% on NTU 120 while modifying fewer than 5% of parameters. Yuan et al. [127] demonstrate successful fine-tuning transfer from NTU to specialized Tai Chi recognition, where pre-trained spatial transformers are frozen and only the classification layers are retrained.

4.4.3. Cross-method comparison and key insight

Environmental adaptation methods form a hierarchy of increasing sophistication. Transfer learning (fine-tuning pre-trained models) provides the simplest approach but assumes moderate domain similarity. Domain adaptation (feature alignment) handles larger distributional shifts but requires target-domain data. Graph heterogeneity alignment addresses structural differences unique to skeleton HAR. The most effective approaches combine strategies: transfer learning for initialization, domain adaptation for distribution alignment, and graph mapping for structural compatibility.

A key finding is that *fixed-topology GCNs transfer better than adaptive ones* across datasets [97], because adaptive topologies overfit to source-domain noise patterns. This suggests a principled trade-off: within-dataset, adaptive topologies improve by discovering action-specific joint correlations; across datasets, fixed anatomical priors provide regularization that prevents encoding environment-specific artifacts. Practitioners should use adaptive graphs only when training and deployment environments are well-matched.

4.4.4. Open challenges

Current domain adaptation methods address pairwise source-target transfer, but real-world deployment involves adaptation to continuously varying conditions. Continual adaptation which updates models incrementally without catastrophic forgetting of previously learned environments remains largely unaddressed; current methods lose 15–20% accuracy on original domains after adaptation [37]. Multi-source transfer, leveraging multiple diverse training environments simultaneously,

could improve robustness but requires investigation into optimal source selection and knowledge-sharing strategies.

4.5. Learning under annotation scarcity

4.5.1. Problem definition and evidence

Annotation costs limit the scalable deployment of skeleton-based HAR, particularly in specialized domains requiring expert knowledge. Medical rehabilitation assessment demands clinician annotations with knowledge of movement disorders; sports analytics requires coaches' understanding of technique subtleties; industrial safety monitoring needs domain-specific action taxonomies. These costs are prohibitive at scale: annotating a clinical rehabilitation sequence may require 15–30min of specialist time, versus seconds for crowdsourced labeling of daily activities.

The performance cost of label scarcity is quantifiable. For purely supervised training on NTU RGB + D, reducing the label budget from 100% to 10% degrades accuracy by 10–15 percentage points. Self-supervised pre-training closes most of this gap in recent benchmarks (see the SSL results later in this subsection), so the 10–15 pp figure should be read as the baseline that label-efficient strategies aim to recover rather than as an unavoidable cost.

4.5.2. Method analysis across paradigms

Self-supervised pre-training achieves near-supervised performance. SSL methods show the largest label efficiency gains. MaskCLR [2] achieves 93.9% on NTU60-XSub under linear evaluation, matching supervised state-of-the-art while requiring zero labels during pre-training. Its Attention-Guided Probabilistic Masking dynamically occludes high-attention joints, forcing models to exploit underutilized motion patterns. S-JEPA [1] predicts high-level latent representations through spectral contrastive loss, reaching 85.3% on NTU60. SCD-Net [112] disentangles spatial and temporal features for complementary pre-training tasks (74.2% on PKU-MMD II). The progression from 3s-HYSP (79.1%) [28] through S-JEPA (85.3%) to MaskCLR (93.9%) shows rapid maturation, with the latest SSL methods eliminating the label-efficiency gap on standard benchmarks.

Semi-supervised learning combines limited labels with structural regularization. When some labels are available, semi-supervised methods leverage structural assumptions about feature-space smoothness. SkeleMoCLR [71] combines momentum contrastive learning with pseudo-labeling and teacher–student networks, matching supervised performance using only 10% labeled data. GRA [37] uses graph representation alignment with dynamic self-training, maintaining accuracy under 40% joint occlusion. SCD-Net (Semi) [112] reaches 82.2% on NTU60-XSub with 10% labels by combining spatiotemporal disentanglement with pseudo-label refinement. PSTL [136] uses Central Spatial Masking and Motion Attention Temporal Masking to simulate occlusions within a Barlow Twins framework, reaching 77.3% with 10% labels.

An important limitation is confirmation bias: pseudo-labeling methods propagate incorrect predictions as training targets. CD-JBF-GCN [99] addresses this through joint-bone feature fusion with self-supervised pose prediction, using complementary representations to cross-validate pseudo-labels. However, a systematic evaluation of confirmation bias severity across different label ratios and action complexity levels has not been conducted.

Active learning selects maximally informative samples. Active learning reduces annotation costs by prioritizing samples whose labels would most reduce model uncertainty. Zhang et al. [133] use multi-scale GNN representations capturing hierarchical relationships from fine-grained joints to coarse-grained limbs, combined with pairwise loss comparison to prioritize high-discrepancy samples. AL-SAR [47] achieves 80% accuracy on NW-UCLA using only 20% labeled data by combining unsupervised representation learning with active sample selection through encoder–decoder architectures and multi-head uncertainty aggregation. These results suggest that strategic labeling can recover 80–90% of supervised performance with 20% of the annotation budget.

GAN-based augmentation expands effective training sets. GANs address label scarcity indirectly by synthesizing additional training samples. AGN [109] integrates action style transfer with active learning, achieving full-dataset performance with only 10% of the original data, the strongest augmentation result in the literature. Kinetic-GAN [21] generates high-quality 34-second sequences using graph convolutions and joint-specific noise modules. However, mode collapse limits generated diversity, and temporal coherence remains challenging: GANs produce realistic individual frames but may violate biomechanical constraints over extended sequences.

Few-shot learning enables rapid adaptation to new action categories. Few-shot methods address a distinct aspect of label scarcity: adaptation to entirely new action categories with minimal examples. STAR [10] decomposes skeletons into topology-based parts using GPT-3.5-generated descriptions, aligning skeleton and semantic spaces for zero-shot recognition. SA-DVAE [48] disentangles semantic-related and semantic-irrelevant features via variational autoencoders. PURLS [138] applies cross-attention to dynamically weight joint features using GPT-3-generated spatial-temporal descriptions, reaching 32.2% on Kinetics-skeleton 200. These results show that language grounding can provide the compositional structure needed for few-shot transfer, though large gaps relative to supervised baselines persist.

4.5.3. Cross-method comparison and key insight

Label-efficient methods form a complementary hierarchy addressing different annotation constraints:

- **Zero labels available**: SSL pre-training (MaskCLR: 93.9%) provides the strongest baseline, approaching supervised performance through self-supervised objectives alone.
- **Small label budget (<20%)**: Semi-supervised learning (SkeleMoCLR at 10%) and active learning (AL-SAR at 20%) efficiently leverage limited annotations. Active learning outperforms random sampling by prioritizing informative examples.
- **New action categories**: Few-shot and zero-shot methods (STAR, PURLS) adapt to unseen classes through language grounding, though accuracy gaps of 30–60 points relative to supervised methods remain.
- **Data augmentation**: GAN-based synthesis (AGN) amplifies small datasets, complementing all other strategies.

The most efficient pipeline combines SSL pre-training with semi-supervised fine-tuning using 10% labels, achieving near-supervised performance. Adding active learning for strategic label selection further reduces the required annotation budget. This combined approach has been independently validated for each component, but not as an integrated pipeline.

4.5.4. Open challenges

The interaction between label efficiency and domain shift is poorly understood. SSL methods achieve near-supervised performance on in-distribution data (NTU), but their label efficiency under domain shift (e.g., pre-training on NTU, fine-tuning with 10% labels on PKU-MMD) has not been systematically evaluated. If SSL representations are as domain-specific as supervised ones, label efficiency gains may not transfer to deployment scenarios. Multi-task self-supervised objectives (MS^2L [54], 3s-AimCLR + + [29]) that jointly optimize motion prediction, contrastive learning, and temporal ordering could provide more robust pre-training, but principled task selection criteria remain absent.

4.6. Open vocabulary and compositional action recognition

4.6.1. Problem definition and evidence

Real-world action vocabularies are open-ended and evolve over time. Training sets cannot enumerate all possible human activities, yet deployed systems must recognize actions absent from the training data. The gap is large: supervised methods reach 93–95% on closed-vocabulary NTU benchmarks, while zero-shot recognition on unseen classes reaches only 32–46%, ranging from PURLS [138] on Kinetics-skeleton 200 at the low end to SA-DVAE [48] on NTU 120 at the high end, with STAR [10] falling in between. This spread corresponds to a 47–63 percentage point gap relative to supervised baselines. Few-shot recognition (5-way 1-shot or 5-shot settings) achieves 32–59% depending on the evaluation protocol, remaining at a similar distance below supervised performance.

The core difficulty is that skeleton sequences lack the semantic richness needed to distinguish actions sharing similar kinematics. “Drinking” and “phone calling” involve nearly identical hand-to-face motions, distinguishable mainly by object context absent in skeletal data. Open vocabulary recognition requires inferring action semantics from motion alone, or grounding skeletal motion in an external knowledge source that supplies the missing context.

4.6.2. Method analysis across paradigms

LLM integration grounds skeletal motion in semantic space. Large Language Models provide the external semantic knowledge needed for open vocabulary recognition. CrossGLG [123] uses GPT-generated textual descriptions for one-shot recognition, enabling classification from textual specifications alone. GAP [117] generates action descriptions (e.g., “person raising right hand overhead”) that guide GCN training through contrastive learning, aligning skeleton and text embeddings. SKI [92] projects skeletons into CLIP’s vision-language embedding space, enabling zero-shot recognition by matching skeleton embeddings to text descriptions of unseen actions. This addresses HAR’s adaptation challenge: generalization to novel action vocabularies without retraining.

LLM-AR [82] takes a different approach: quantizing skeleton sequences into discrete tokens via VQ-VAE, constructing “action sentences” from token sequences, and fine-tuning LLaMA with LoRA adapters. This achieves 95.0% on NTU (the highest reported accuracy) but operates in a closed vocabulary setting. The result shows that LLMs can boost recognition accuracy without directly addressing open vocabulary generalization.

Language-assisted learning transfers semantic structure to skeleton encoders. LA-GCN [120] and VL [11] distill language knowledge into skeleton-specific models, enabling deployment without LLM inference overhead. LA-GCN transfers language representations to GCN-based recognition, achieving 93.5% on NTU 60 and 97.6% on NW-UCLA with standard GCN inference latency. VL uses progressively distilled cross-modal knowledge to establish joint action concept spaces, reducing noise through intra-modal similarity and inter-modal consistency. These approaches retain semantic benefits while eliminating the 200–500 ms LLM latency penalty.

Disentangled representations support compositional generalization. SA-DVAE [48] disentangles features into semantic-related and semantic-irrelevant components via modality-specific variational autoencoders with adversarial total correlation penalties. By isolating the semantic component, the model can recombine learned motion primitives for unseen action categories. STAR [10] achieves 59.4% harmonic mean in generalized zero-shot learning on NTU by decomposing skeletons into topology-based parts with dual prompts for intra-class compactness and inter-class separability. PURLS [138] introduces part-aware unified representation integrating language and skeleton modalities through adaptive partitioning, reaching 32.2% on Kinetics-skeleton 200.

4.6.3. Cross-method comparison and key insight

Open vocabulary methods present a core trade-off between semantic capability and deployment feasibility. Full LLM integration (LLM-AR, SKI) provides the richest semantic grounding but requires 200–500ms inference, which is prohibitive for real-time applications. Language-assisted learning (LA-GCN, VL) distills semantic knowledge into efficient models, retaining most accuracy gains at standard latency. Disentangled representations (SA-DVAE, STAR) enable compositional generalization from skeletons alone but trail language-grounded methods in accuracy.

A key finding is that textual descriptions function as a form of *structured annotation* rather than true zero-shot learning. Methods require curated action descriptions as input (whether human-authored or GPT-generated), and performance scales with description quality. Vague descriptions (“person moves arms”) provide minimal signal, while detailed specifications enable fine-grained discrimination. This dependency creates a secondary annotation bottleneck that partially undermines the promise of label-free recognition. Furthermore, LLMs trained on Western-centric text corpora exhibit cultural biases: actions common in underrepresented cultures (traditional dances, ceremonial gestures) lack rich textual descriptions, degrading cross-modal alignment in global deployment contexts.

4.6.4. Open challenges

Lightweight zero-shot models (<1B parameters) meeting real-time latency budgets (<100ms) do not exist. Distilling knowledge from large LLMs into compact skeleton encoders achieves partial success (LA-GCN) but at the cost of reduced zero-shot flexibility. Whether skeleton-specific foundation models pre-trained on massive unlabeled motion datasets can match LLM-grounded performance through skeleton-native representations, eliminating the language dependency, remains a central open question.

4.7. Computational efficiency for edge deployment

4.7.1. Problem definition and evidence

Real-world deployment imposes strict latency and resource constraints. Virtual reality requires sub-50ms response times for immersive interaction; industrial safety monitoring demands sub-100ms recognition for real-time intervention; mobile health applications must operate within the power budgets of wearable devices. Current top-performing methods violate these constraints: GCN inference requires 130ms (ST-GCN-level methods [76]), transformers exceed 200ms, and LLM-integrated approaches add 200–500ms latency [82].

The efficiency challenge is compounded by a hardware mismatch. Skeleton data is sparse by nature: 25 joints with 3D coordinates yield 75 floating-point values per frame. Yet modern accelerators (GPUs, NPUs) are optimized for dense tensor operations. Graph convolutions involve sparse-dense matrix multiplications reaching only 60–70% hardware utilization [90], translating theoretical FLOPs advantages into modest wall-clock speedups.

4.7.2. Method analysis across paradigms

CNN architectures dominate the latency-accuracy pareto frontier. Pruned CNN models achieve the best latency-accuracy trade-off for real-time deployment. Noor et al. [76] show 4.10ms inference with 98.35% accuracy through depthwise separable convolutions and structured pruning on a ResNet18 backbone, meeting VR latency requirements while maintaining competitive accuracy. PoseC3D [26] achieves 93.7% on NTU with moderate computational cost, benefiting from mature GPU optimization for regular convolution operations. The CNN advantage stems from hardware alignment: regular grid operations fully utilize tensor core parallelism, achieving near-100% hardware efficiency versus GCNs’ 60–70%.

Efficient GCN designs reduce but do not eliminate the latency gap. EchoGCN [81] achieves 0.7G FLOPs, an order of magnitude below typical GCNs (15–20G), through echo-based graph convolution that

reuses intermediate features, reaching 92.8% on NTU 60 cross-subject. BlockGCN [137] cuts parameters by 40% via static-dynamic block encoding while maintaining 91.5% on NTU 120. SHoTGCN [66] uses re-parameterized convolutions for shoulder-torso-hip modeling, reaching 93.6% with improved efficiency. However, even optimized GCNs face the sparse operation bottleneck: 0.7G FLOPs on sparse graph operations translate to considerably longer wall-clock time than 0.7G FLOPs of dense convolutions.

Transformer efficiency optimizations trade complexity for approximation. SkateFormer [23] partitions skeletons into four anatomical groups with hierarchical attention, reducing complexity by 48× while achieving 93.5% accuracy. This partition strategy effectively approximates full self-attention by assuming that within-group interactions dominate, a reasonable assumption for many actions but potentially limiting for activities requiring full-body coordination. Despite optimizations, transformer latency remains above 100ms for 300-frame sequences, restricting deployment to offline analysis scenarios.

Knowledge distillation transfers capabilities to compact models. T2SPNet [140] transfers temporal knowledge from high-quality teachers to compact students, reducing computational costs by 20% while maintaining accuracy. Cross-modal distillation [96] trains skeleton-processing students from RGB-trained teachers, achieving multimodal performance at skeleton-only inference costs. MKE-GCN [126] uses multiple specialized teachers to supervise a unified compact student. These approaches decouple training complexity from inference efficiency, enabling the deployment of knowledge-rich models within resource constraints.

4.7.3. Cross-method comparison and key insight

The latency-accuracy landscape separates into three deployment tiers:

- **Real-time (<50ms)**: CNN-only approaches. Pruned ResNets [76] at 4ms; PoseC3D variants at 10–30ms. Suitable for VR/AR and robotics.
- **Near-real-time (50–200ms)**: Efficient GCNs (EchoGCN, BlockGCN) and distilled models. Suitable for surveillance, and industrial monitoring.
- **Offline (>200ms)**: Transformers and LLM-integrated methods, suitable for forensic and sports video analysis.

The practical implication is that method selection is constrained first by the latency budget and only then by accuracy. A practitioner needing sub-50ms response has no GCN, transformer, or LLM options regardless of their accuracy. Benchmark-driven research often overlooks this constraint, reporting accuracy without latency context.

4.7.4. Open challenges

Hardware-software co-design exploiting skeleton data sparsity represents the most promising direction. Neuromorphic computing (event-driven processing exemplified by DHP19 [5]) achieves sub-millisecond latency by processing only changed joint positions. NPU-optimized sparse graph operations could eliminate the hardware utilization gap. Multi-objective neural architecture search discovering skeleton-specific architectures that jointly optimize accuracy, latency, and power consumption remains unexplored. Quantization-aware training, which has achieved 4–8× speedups for image classification, has not been systematically evaluated for skeleton HAR, where the already-sparse input may be more sensitive to precision reduction.

4.8. Cross-cutting synthesis

4.8.1. Method–challenge suitability matrix

Table 4 maps method families onto generalization challenges, synthesizing the analyses from Sections 4.2–4.7. We assign each cell based on the quantity and strength of empirical results on the relevant benchmarks rather than on intuition: **Strong** (S) requires at least two peer-reviewed methods within 5 percentage points of the best reported result (or, for efficiency, methods meeting the latency tier of Section 4.7); **Moderate** (M) requires a single paper within 15 percentage points of the best, or positive results restricted to one benchmark; **Weak** (W) is assigned when reported results lag the best by more than 15 percentage points or the family suffers an architectural mismatch; a dash (–) indicates that no published work from that family has been evaluated on that challenge. Each cited method substantiates the rating in its row.

The matrix shows several patterns. First, most method families achieve at most two Strong ratings; self-supervised pre-training is the only paradigm spanning three (noise, cross-subject, and annotation scarcity), and the field still lacks single-paradigm coverage of more than half of the challenges, motivating the integrated pipelines analyzed in Section 4.8. Second, SSL and CNN emerge as the most versatile paradigms: SSL addresses noise robustness, cross-subject generalization, and label scarcity simultaneously [1,2,71], while CNNs combine noise robustness [26] with computational efficiency [76]. Third, notable evaluation gaps remain: LLM-based methods have not been evaluated under sensor noise, cross-subject demographics, or viewpoint shift, and meet only the offline efficiency tier (≥200 ms latency [82]); domain adaptation has not been evaluated for noise robustness, label scarcity, open vocabulary, or computational efficiency; and no method achieves a Strong rating jointly on open vocabulary and sub-100 ms inference. These ratings reflect the current balance of published evidence, much of it obtained under heterogeneous datasets and protocols, rather than definitive rankings of method families. For cross-dataset transfer, demographic shift, and open-vocabulary recognition in particular, controlled comparative studies remain scarce, so the corresponding ratings carry lower confidence and may shift as standardized benchmarks appear.

4.8.2. Challenge co-occurrence in deployment scenarios

Real deployments face multiple challenges simultaneously. Table 5 maps four representative deployment scenarios to their dominant generalization challenges and recommends method combinations based on the suitability analysis.

These recommendations come with a caveat. No existing method or pipeline has been validated against simultaneous challenges; each recommendation combines components that were validated independently, and interaction effects remain unknown. For example, a CNN reaching 47.7% noise robustness on Kinetics and an SSL method reaching 93.9% label efficiency on NTU may not combine additively when both challenges co-occur. Systematic evaluation of method combinations under multi-challenge conditions is therefore a prerequisite for deployment-ready solutions, and we list it among the short-term research priorities in Section 6.

4.8.3. Research effort vs. challenge severity

The contrast between challenge severity (Table 3) and research coverage indicates a systematic mismatch. Cross-subject and cross-view generalization (the smallest gaps, 3–7 pp) dominate evaluation protocols: every paper in Tables 10–12 reports NTU cross-subject and/or cross-view results. Sensor noise robustness (the largest gap, 46–60 pp) is evaluated by fewer than 15% of methods (those reporting Kinetics results). Open vocabulary recognition and demographic shift, each exceeding 35 pp gaps, are addressed by fewer than 10 papers in our survey.

This mismatch has structural causes. NTU RGB+D provides standardized cross-subject and cross-view protocols that enable convenient benchmarking; no comparably standardized protocols exist for noise robustness (Kinetics lacks official skeleton annotations), demographic generalization (no cross-demographic benchmark exists), or domain adaptation (no paired source-target skeleton datasets with matched action vocabularies). Addressing the most impactful challenges requires evaluation infrastructure commensurate with the problems' severity.

Table 4
Method–challenge suitability matrix for skeleton-based HAR. Cells apply the rating criteria of Section 4.8: **S** = Strong, **M** = Moderate, **W** = Weak, **–** = not evaluated. Each row label links to the per-paradigm analysis with citations.

Challenge (evidence in)	RNN	CNN	GCN	Trans.	LLM	SSL	DA	KD
C1 Sensor noise (Section 4.2)	W	**S**	W	M	–	**S**	–	M
C2 Cross-subject / demographic (Section 4.3)	M	M	**S**	M	–	**S**	**S**	M
C3 Viewpoint / environment (Section 4.4)	M	M	M	M	–	M	**S**	M
C4 Annotation scarcity (Section 4.5)	W	W	M	W	M	**S**	–	M
C5 Open vocabulary (Section 4.6)	W	W	W	W	**S**	M	–	M
C6 Computational efficiency (Section 4.7)	M	**S**	M	W	W	–	–	**S**

Table 5
Representative deployment scenarios, their dominant generalization challenges, and recommended pipeline compositions. Per-paradigm evidence and citations are deferred to Table 4 and Sections 4.2–4.7.

Scenario	Dominant challenges	Recommended pipeline
Hospital monitoring	C1 + C2 + C4 + C6	CNN heatmap backbone (C1, C6); SSL pre-training to recover accuracy from limited expert annotations (C4); DA across healthy and patient cohorts (C2).
Smart home (elderly)	C2 + C5 + privacy	GCN with anatomical prior for sample efficiency (C2); LLM-assisted recognition for open-ended household actions (C5); federated or differentially private training for in-home data residency.
Industrial safety	C1 + C3 + C4 + C6	Pruned CNN for sub-50 ms noise-robust inference (C1, C6); GAN-based occlusion completion and action-style augmentation for rare events (C4); DA across factory cameras and viewpoints (C3).
Sports analytics	C2 + C4 + fine-grained motion	Transformer backbone for long-range, fine-grained joint dependencies; SSL pre-training for label efficiency on specialist annotations (C4); multi-stream and ensemble fusion for the final accuracy margin (C2).

5. Datasets, evaluation protocols, and performance analysis

The generalization gaps quantified in Section 4 are not only measurement artifacts; they are direct consequences of how current benchmarks are built. NTU RGB + D provides clean Kinect skeletons within a narrow demographic, so the most severe challenges identified in Section 4 (noise, demographic shift, open vocabulary) remain invisible to standard evaluation protocols, while the least severe (cross-subject and cross-view) dominate leaderboards. In this sense, the design choices behind existing datasets are the root cause of the generalization gaps documented in Section 4. This section examines how dataset scale, sensor type, participant recruitment, and splitting strategies determine which challenges a benchmark can measure and which it silently hides. We then assess whether current evaluation protocols can support deployment-ready method selection, and we identify the benchmark gaps that must be closed before the field's measured progress translates into real-world reliability.

5.1. Dataset overview: what challenges does each benchmark test?

Tables 6–8 document dataset characteristics. Rather than enumerating these datasets chronologically, we assess each through the lens of the generalization challenges it enables or fails to test.

5.1.1. The controlled environment trap

Small-scale datasets (Table 6) share a common limitation: actors perform predefined actions on command in laboratory settings. This controlled paradigm enables reproducible evaluation of cross-subject (Challenge C2) and cross-view (Challenge C3) generalization within narrow bounds, but systematically excludes the noise (C1), demographic diversity (C2), environmental variability (C3), and open-ended action vocabularies (C5) inherent to real-world deployment. The resulting 30–40% degradation in field deployments [17,26] reflects not insufficient scale but a distributional mismatch between controlled performance and naturalistic behavior.

5.1.2. Multimodal complexity without demographic diversity

Medium-scale datasets (Table 7) add multimodal sensors and varied environments but inherit demographic homogeneity: subjects are predominantly university students aged 18–35. This creates an illusion of robustness through sensor diversity while maintaining population uniformity. Methods may learn correlations between young adult biomechanics and action labels rather than generalizable motion patterns, consistent with the 20–40 percentage point degradation reported across elderly-inclusive benchmarks [20,39]. Multimodal sensing cannot compensate for demographic sampling bias; Challenge C2 remains systematically untested.

5.1.3. Scale without challenge coverage

Large-scale datasets (Table 8) achieve large sample counts but vary widely in the challenges they test. NTU RGB + D [61,85] provides clean Kinect-captured skeletons with standardized cross-subject and cross-view protocols, ideal for C2 (cross-subject) and C3 (viewpoint) within its demographic range but unable to test C1 (noise) due to clean capture. Kinetics-Skeleton [122] tests C1 through pose-estimated noisy skeletons and C5 (open vocabulary) through its 400-class vocabulary, but lacks cross-subject protocols due to unknown subject identities. ETRI-Activity3D [39] partially tests C2 (demographics) with 100 subjects but remains controlled. AddBiomechanics [111] offers biomechanically rich data with 273 subjects and 24M frames, a promising direction for physics-informed analysis, but its 9 activity classes limit applicability.

The underlying tension persists: *no single dataset tests more than two generalization challenges simultaneously*. NTU tests cross-subject and cross-view; Kinetics tests noise and vocabulary breadth; ETRI tests demographic breadth within controlled settings. Bridging this gap requires new benchmarks combining population diversity, naturalistic capture, noise variation, and sufficient scale—a dataset design challenge that mirrors the method design challenge identified in Section 4.8.

5.2. Evaluation metrics and their limitations

Standard metrics capture limited aspects of model behavior and systematically obscure generalization failures.

Top-1 accuracy, the dominant metric, conceals class-imbalanced performance. Methods achieving 93% on NTU RGB + D may classify frequent actions correctly while failing on rare but important categories (e.g., falling, abnormal gait). **Top-k accuracy** allows hedging rather

Table 6
Small-scale skeleton-based HAR datasets and the generalization challenges they enable.

Dataset	# Class	# Subject	# Sample	# Joint	Source	Characteristics*				
						MV	SA	IA	DA	AA
HDM05 [73]	130	5	2337	31	Motion capture system		✓		✓	
MSR Action3D [50]	20	10	567	20	Microsoft Kinect		✓		✓	
UTKinect [116]	10	10	200	20	Microsoft Kinect		✓		✓	
MSRDaily Activity3D [103]	16	10	320	20	Microsoft Kinect		✓	✓	✓	
CAD-60 [95]	12	4	60	15	Microsoft Kinect		✓		✓	
CAD-120 [45]	10	4	120	15	Microsoft Kinect		✓	✓	✓	
UCFKinect [27]	16	16	1280	15	Microsoft Kinect		✓		✓	
Berkeley MHAD [77]	11	12	660	35	Motion capture system	✓	✓		✓	
SBU [101]	8	7	300	15	Microsoft Kinect		✓	✓	✓	
IAS-Lab [74]	15	12	540	27	Microsoft Kinect		✓		✓	
NW-UCLA [102]	10	10	1494	20	Microsoft Kinect	✓	✓		✓	
UTD-MHAD [7]	27	8	861	20	Microsoft Kinect		✓		✓	
DHP19 [5]	33	17	561	13	DVS event cameras	✓	✓		✓	

* MV: Multi-view, SA: Single-person Activities, IA: Interactive Activities, DA: Daily Activities, AA: Abnormal Activities.

Table 7
Medium-scale skeleton-based HAR datasets.

Dataset	# Class	# Subject	# Sample	# Joint	Source	Characteristics*				
						MV	SA	IA	DA	AA
JHMDB [40]	21	–	31k	13	Videos	✓	✓		✓	
PKU-MMD [56]	51	66	1076	25	Microsoft Kinect	✓	✓	✓	✓	✓
Drive&Act [72]	83	15	–	13	Video	✓	✓		✓	✓
MMAct [43]	37	20	36k	17	Video	✓	✓	✓	✓	
TSU [19]	31	18	16k	13	Microsoft Kinect	✓	✓		✓	
EV-Action [105]	20	70	7000	25	Vicon camera & Kinect	✓	✓		✓	
MM-Fi [125]	27	40	1080	17	Customized sensor platform		✓		✓	
RT-Pose [33]	6	10	72k	15	4D Radar Tensor, LiDAR, RGB	✓			✓	

* MV: Multi-view, SA: Single-person Activities, IA: Interactive Activities, DA: Daily Activities, AA: Abnormal Activities.

Table 8
Large-scale skeleton-based HAR datasets.

Dataset	# Class	# Subject	# Sample	# Joint	Source	Characteristics*				
						MV	SA	IA	DA	AA
NTU RGB+D 60 [85]	60	40	56k	25	Microsoft Kinect	✓	✓	✓	✓	✓
NTU RGB+D 120 [61]	120	106	114k	25	Microsoft Kinect	✓	✓	✓	✓	✓
Kinetics-Skeleton [122]	400	400	300k	18	Video		✓	✓	✓	✓
ETRI-Activity3D [39]	55	100	112k	25	Microsoft Kinect		✓		✓	✓
UAV-Human [49]	155	119	67k	17	UAV platform (Kinect)		✓		✓	
AddBiomechanics [111]	9	273	24M	37	Optical Motion Capture, Force Plates	✓			✓	

* MV: Multi-view, SA: Single-person Activities, IA: Interactive Activities, DA: Daily Activities, AA: Abnormal Activities.

than discriminative commitment. Neither metric assesses *robustness* (resilience to noise injection), *calibration* (whether 90% confidence predictions are correct 90% of the time), or *cross-dataset transfer* (performance in unseen environments).

Connecting to the challenge framework: robust evaluation of Challenge C1 (noise) requires reporting accuracy under controlled noise injection at multiple severity levels. Challenge C2 (demographics) requires per-demographic breakdowns. Challenge C3 (environment) requires mandatory cross-dataset reporting. Challenge C4 (labels) requires learning curves across label ratios. Challenge C6 (efficiency) requires FLOPs, latency, and parameter counts alongside accuracy. Current benchmarks report none of these systematically.

5.3. Splitting strategies and protocol limitations

Splitting strategies determine which generalization challenges a benchmark can test, but protocol design choices introduce systematic biases. Table 9 summarizes common approaches.

Current protocols each test a single factor while holding others fixed. **Cross-subject splitting** on NTU (20/20 subjects) tests subject variation within a homogeneous demographic, producing 3–7 pp gaps. However, when all subjects perform identical actions in identical environments, models exploit dataset-specific cues rather than learning subject-invariant patterns, explaining the 33-point degradation when transferring to PKU-MMD [36]. **Cross-view** and **cross-setup** protocols test viewpoint and environmental robustness under controlled conditions that poorly reflect uncontrolled deployment. True deployment robustness demands evaluation under *joint* variation of demographics, environments, sensors, and action contexts, a compound challenge that no current protocol assesses.

5.4. Performance analysis through the generalization lens

Tables 10–12 document method performance across major benchmarks. We analyze these results through the challenge framework rather than as isolated accuracy comparisons.

5.4.1. Benchmark saturation and the challenge mismatch

Table 10 shows saturation on NTU RGB+D 60: over 10 methods achieve 93–95% cross-subject, with incremental differences (0.5–1%) within experimental noise margins (±0.3–0.5% across runs). Cross-view shows similar saturation: 8 methods exceed 97%. NW-UCLA: 9 methods

Table 9
Dataset splitting protocols and the generalization challenges they assess.

Protocol	Representative Datasets	Strengths	Limitations	Challenge
Random Split	MSR Action3D [50], UTKinect [116]	Baseline benchmarking	High data leakage risk; poor generalization assessment	–
Cross-Subject	NTU RGB+D [85], PKU-MMD [56], ETRI [39]	Tests subject-invariance	Within homogeneous demographics; sensitive to inter-class similarity	C2
Cross-View	NTU RGB+D [85], Human3.6M [38], UAV-Human [49]	Tests viewpoint robustness	Fixed camera positions; does not test environmental variation	C3
Cross-Setup	NTU RGB+D 120 [61], MMAct [43]	Tests environmental changes	Requires collection metadata; controlled variation only	C3
Task-Specific	Kinetics [122], AddBiomechanics [111]	Application-aligned evaluation	Lacks standardization	Varies

Table 10
Performance of representative skeleton-based HAR models across benchmarks.

Method	Year	NTU RGB+D 60		NTU RGB+D 120		Kinetics	NW-UCLA
		C-Sub	C-View	C-Sub	C-Set	Top-1	Acc
Recurrent Neural Networks (RNN)							
ST-LSTM [62]	2016	61.7	75.5	55.7	57.9	–	–
IndRNN [134]	2018	81.8	88.0	–	–	–	–
EleAtt-RNN [132]	2020	80.7	88.4	–	–	–	90.7
Convolutional Neural Networks (CNN)							
Clips+CNN+MTLN [42]	2017	79.6	84.8	58.4	57.9	–	–
3SCNN [52]	2019	88.6	93.7	–	–	–	–
PoseC3D [26]	2022	93.7	96.6	86.9	90.3	**47.7**	–
MSST [15]	2023	89.6	95.3	85.3	86.0	–	95.3
Graph Convolutional Networks (GCN)							
ST-GCN [122]	2018	81.5	88.3	79.0	–	33.4	–
InfoGCN [16]	2021	93.0	97.1	88.5	89.7	–	97.0
CTR-GCN [13]	2021	92.4	96.8	88.9	90.6	–	96.5
BlockGCN [137]	2024	93.1	97.0	90.3	91.5	–	96.9
EchoGCN [81]	2025	92.8	96.7	89.7	90.9	–	96.7
SD-GCN [41]	2025	93.1	97.1	89.2	91.2	–	97.0
SHoTGCN [66]	2025	93.6	97.5	**91.2**	**92.9**	–	96.3
SCA-GCN [58]	2025	89.8	96.0	–	–	37.3	–
Chen-HSIC [12]	2025	93.7	97.3	90.6	91.7	–	97.2
MAR-GCN [30]	2026	92.3	96.5	87.6	90.3	–	96.6
Large Language Model (LLM) Based							
LLM-AR [82]	2024	**95.0**	**98.4**	88.7	91.5	–	–
LA-GCN [120]	2025	93.5	97.2	89.7	90.9	–	97.6
Transformer / Attention Based							
ST-TR-agcn [80]	2020	89.9	96.1	82.7	84.7	37.4	–
Hyperformer [135]	2022	92.9	96.5	89.9	91.3	–	96.7
FreqMixFormer [114]	2024	93.6	97.4	90.5	91.9	–	97.7
SkateFormer [23]	2024	93.5	97.8	89.8	91.4	–	**98.3**
LG-STFormer [67]	2025	92.8	96.7	88.6	90.4	–	97.0

Results collected from original papers. Dashes indicate unreported metrics.

achieve 96–98.3%. These plateaus indicate that NTU RGB+D has exhausted its utility for discriminating method quality on Challenges C2 (cross-subject) and C3 (cross-view), the very challenges where gaps are smallest.

In contrast, Kinetics-Skeleton, which tests Challenge C1 (noise), shows large performance differences: PoseC3D achieves 47.7% while GCN methods reach only 33.4–37.3%. Self-supervised MaskCLR achieves 44.7%. This 14-point spread on Kinetics provides far more discriminative power than the 2-point spread on NTU, yet fewer than 15% of methods report Kinetics results. The field optimizes for the benchmark where differentiation is exhausted while neglecting the benchmark where genuine performance differences exist.

5.4.2. Cross-dataset gaps confirm artifact learning

Domain adaptation results in Table 11 quantify Challenge C3 (environment) and confirm the artifact learning hypothesis. NTU-to-NW-UCLA transfer yields 82.5–86.3% [97,121] versus 96–98% target-trained. NTU-to-PKU transfer achieves only 51.6% [36] versus 85.1% target-trained. These 10–33 point gaps persist despite shared action vocabularies, indicating that models learn camera-specific noise patterns, background regularities, and demographic-specific motion styles rather than transferable action representations.

5.4.3. The missing dimensions

Performance tables report only accuracy. Important dimensions absent from current evaluation include:

- **Computational cost**: EchoGCN achieves 0.7G FLOPs versus typical GCNs at 15–20G, yet accuracy differs by only 0.2–0.8%. Latency comparisons (4ms CNN vs. 200 + ms Transformer vs. 500ms LLM) are unreported.
- **Noise robustness**: Methods achieving 93% on clean NTU may degrade by 15–30% under joint occlusion or Gaussian noise [79], but robustness metrics are absent from standard reporting.
- **Class-wise breakdown**: Aggregate accuracy may mask failure on rare but safety-relevant actions (falling in healthcare, equipment misuse in industry).

Table 11
Performance of advanced skeleton-based HAR methods (SSL, semi-supervised, transfer, few-shot).

Method	Year	NTU RGB+D 60		NTU RGB+D 120		Kinetics	NW-UCLA
		C-Sub	C-View	C-Sub	C-Set	Top-1	Acc
Generative Adversarial Networks (GAN)							
VN-GAN [78,83]	2023	92.5	87.6	89.4	87.6	–	**95.9**
Self-Supervised Learning (Linear Evaluation)							
3s-HYSP [28]	2023	79.1	85.2	64.5	67.3	–	–
S-JEPA [1]	2024	85.3	89.8	79.6	79.9	–	–
SCD-Net [112]	2024	86.6	91.7	76.9	80.1	–	–
MaskCLR [2]	2024	**93.9**	**97.3**	**87.4**	**89.5**	**44.7**	–
IGM [55]	2024	91.2	86.2	81.4	80.0	–	–
SG-CLR [68]	2025	88.2	94.1	83.0	83.6	–	–
Semi-Supervised Learning (10% Labeled Data unless noted)							
CD-JBF-GCN [99]	2023	78.0	71.7	–	–	36.5	–
PSTL [136]	2023	77.3	81.8	69.2	70.3	–	–
SCD-Net (Semi) [112]	2024	**82.2**	**85.8**	–	–	–	–
GRA [37]	2024	78.5	82.5	–	–	–	83.5 (30%)
SkeleMoCLR [71]	2025	78.4	81.8	–	–	–	81.1 (30%)
SG-CLR [68]	2025	81.3	84.6	–	–	–	–
Transfer Learning							
Liao et al. [53]	2022	–	–	–	–	–	86.9
SkeleTR [25]	2023	**94.8**	**97.7**	87.8	88.3	–	–
ViA [124]	2024	78.1	85.8	69.2	66.9	–	–
Skeleton-Prompt [70]	2026	93.6	96.8	**90.2**	**91.0**	–	–
Zero-Shot Learning (Unseen-Class Splits)							
SynSE [31]	2021	75.8 (5u)	–	38.7 (24u)	–	–	–
STAR [10]	2024	45.1 (12u)	42.5 (12u)	44.3 (24u)	44.1 (24u)	–	–
SA-DVAE [48]	2024	41.4 (12u)	–	**46.1** (24u)	–	–	–

For semi-supervised methods, results at 10% labels unless marked. For zero-shot rows, 'u' denotes the number of unseen classes (e.g., 12u uses the 48/12 split on NTU 60 and the 96/24 split on NTU 120). Domain adaptation methods use different protocols (see Section 4.4).

Table 12
Performance of hybrid skeleton-based HAR methods (fusion, ensemble, multi-task, distillation).

Method	Year	NTU RGB+D 60		NTU RGB+D 120		Kinetics	NW-UCLA
		C-Sub	C-View	C-Sub	C-Set	Top-1	Acc
Multi-Stream Fusion							
CNN-LSTM [46]	2017	82.9	90.1	–	–	–	–
MS-AAGCN [88]	2020	90.0	96.2	–	–	37.8	–
MMCL [60]	2024	93.5	**97.4**	**90.3**	**91.7**	–	**97.5**
Model Ensemble							
STF-Net [113]	2023	91.1	96.5	86.5	88.2	36.1	–
ML-STGNet [139]	2023	91.9	96.2	88.6	90.0	38.9	–
Hulk (ViT-L) [110]	2023	**94.3**	–	–	–	–	–
FRF-GCN [129]	2024	91.3	96.5	87.1	88.4	37.9	–
SA-TDGFormer [8]	2025	92.7	96.8	87.8	88.9	**39.0**	–
Multi-Task Learning							
MS^2L [54]	2020	78.6	–	–	–	–	86.8
3s-AimCLR++ [29]	2024	87.1	93.0	82.5	83.2	–	–
2D-SCHAR [106]	2024	–	–	–	–	37.1	–
5PM-STGCN [6]	2024	88.7	95.2	–	–	36.4	–
DSDC-GCN [141]	2024	93.0	97.1	89.9	90.6	38.6	97.4
Knowledge Distillation							
Cross-modal KD [96]	2019	79.5	–	–	–	–	–
MKE-GCN [126]	2022	92.5	96.9	89.7	91.1	–	–
ACFL [108]	2022	92.5	97.1	89.7	90.9	–	–
T2SPNet [140]	2023	90.1	95.5	84.0	85.6	37.5	–
VL [11]	2024	88.3	92.8	82.5	84.3	–	–
Part-level KD [57]	2024	90.1	91.0	–	–	–	–
LA-GCN [120]	2025	93.5	97.2	**90.7**	**91.8**	–	**97.6**

Some results rounded to one decimal place for consistency.

- **Calibration**: Whether high-confidence predictions are reliable (key for safety) is never assessed.

A multi-dimensional evaluation framework reporting accuracy alongside FLOPs, latency, noise robustness, cross-dataset gaps, and calibration error would better support method selection for specific deployment challenges.

6. Future research directions

The generalization analysis in Sections 4 and 5 shows ongoing gaps between benchmark performance and deployment readiness. This section organizes future research around the unresolved challenges, structured as a three-tier roadmap: short-term refinements addressing immediate deployment blockers, medium-term paradigm shifts requiring methodological breakthroughs, and long-term grand challenges demanding major advances.

6.1. Short-term: closing evaluation and efficiency gaps

6.1.1. Benchmark reform for challenge-oriented evaluation

Addresses: All challenges (C1–C6).

Benchmark saturation on NTU RGB+D (10 + methods within 93–95%) signals that the field's primary evaluation can no longer discriminate method quality. The largest generalization gaps (sensor noise, 46–60 pp; open vocabulary, 47–63 pp) receive the fewest evaluations (Table 3). Reforming evaluation requires:

- **Robustness evaluation suites**: Systematic noise injection at multiple severity levels, standardizing the controlled experiments that currently appear only in individual papers [79]. The SKELTER framework provides foundations requiring community adoption.
- **Cross-dataset leaderboards**: Mandatory transfer reporting alongside within-dataset accuracy. A standardized NTU→PKU-MMD and NTU→NW-UCLA protocol would expose artifact learning.
- **Multi-dimensional scorecards**: Reporting accuracy alongside FLOPs, latency, robustness under noise, cross-dataset gap, and calibration error, applying the matrix in Table 4 to each new method.
- **Cross-demographic benchmarks**: A controlled dataset combining matched action vocabularies across age groups (young adult, middle-aged, elderly), body types, and ability levels. ETRI-Activity3D [39] provides partial coverage; a comprehensive cross-demographic benchmark is required for Challenge C2.
- **Multi-challenge evaluation protocols**: Standardized benchmarks that simultaneously test combinations of noise, demographic variation, and label scarcity, validating whether independently effective methods retain their advantages when challenges co-occur (cf. the caveat in Section 4.8).

6.1.2. Efficiency-accuracy optimization for edge deployment

Addresses: Challenge C6 (computational efficiency).

Current top-performing methods violate real-time constraints: GCNs require 130 ms, transformers exceed 200 ms, and LLMs add 200–500 ms latency. CNNs are the only paradigm meeting sub-50 ms budgets (Section 4.7). Closing the gap for other paradigms requires:

- **Skeleton-specific hardware co-design**: Exploiting the inherent sparsity of skeleton data (25 joints × 3 coordinates = 75 values per frame) through neuromorphic computing [5], NPU-optimized sparse operations, and custom accelerators for graph message passing.
- **Quantization-aware training**: Systematically evaluating INT8/INT4 quantization for skeleton models, where the already-sparse input may be more sensitive to precision reduction than image data.
- **Multi-objective NAS**: Discovering skeleton-specific architectures jointly optimizing accuracy, latency, and power consumption rather than adapting image-classification architectures.

6.2. Medium-term: addressing the largest gaps

6.2.1. Physics-informed noise robustness

Addresses: Challenges C1 (noise), C2 (demographics).

Current methods treat skeleton sequences as abstract signals, permitting physically implausible joint configurations. Physics-informed approaches could reject biomechanically impossible poses before they corrupt downstream recognition, addressing the noise challenge at its source. Directions include:

- **Differentiable physics engines**: Integrating musculoskeletal simulation (e.g., MuJoCo [98]) as differentiable layers that constrain model outputs to physically plausible motion.
- **Confidence-weighted message passing**: Integrating pose estimator confidence scores into GCN message passing, weighting aggregation by joint reliability rather than treating all inputs equally.
- **Causal skeleton models**: Discovering joint dependency structures via directed acyclic graphs, separating biomechanical causation (elbow rotation causing wrist displacement) from statistical correlation, improving robustness to distribution shifts.

AddBiomechanics [111] demonstrates the value of physics-informed data, but current constraints reduce recognition accuracy by 15–20% due to a bias toward average poses. Achieving physics-informed regularization without an accuracy penalty remains an open challenge.

6.2.2. Generative motion understanding for open vocabulary

Addresses: Challenges C5 (open vocabulary), C4 (annotation scarcity).

Discriminative models map sequences to fixed label sets without understanding motion generation processes, limiting generalization to unseen actions. Generative approaches could enable:

- **Motion foundation models**: VAEs learning disentangled latent representations separating action semantics from execution style, enabling zero-shot recognition through compositional generalization.
- **Diffusion-based augmentation**: Generating plausible motion sequences conditioned on textual descriptions or partial observations, addressing both data scarcity and vocabulary expansion.
- **Lightweight language grounding**: Distilling LLM semantic knowledge into models under 1B parameters meeting real-time latency budgets, and bridging the gap between LLM-AR's accuracy (95.0%) and edge deployment requirements (<50ms).

6.3. Long-term: toward motion intelligence

6.3.1. Unified motion foundation models

Addresses: All challenges simultaneously.

HAR remains fragmented across isolated tasks with separate pipelines, while NLP and vision have shown that unified architectures handle diverse tasks through shared representations. Skeleton-based motion intelligence requires:

- **Skeleton-native pre-training**: Training on billions of unlabeled sequences from diverse sources (motion capture, video-estimated, wearable sensors), learning universal representations via self-supervised objectives.
- **Multi-task architectures**: Handling recognition, forecasting, generation, anomaly detection, and quality assessment through a single model, exploiting shared motion understanding.
- **Cross-format adaptation**: Automatically aligning heterogeneous skeleton formats (varying joint counts, coordinate systems, noise characteristics) during pre-training, addressing Challenge C3 at the foundation level.

Key barriers include heterogeneous skeleton sources, immature unified tokenization, and catastrophic forgetting (15–20% accuracy loss on original domains after adaptation [37]). Success would mean matching specialized models on 5 + tasks and 60–70% zero-shot generalization (vs. current 35–40%).

6.3.2. Embodied AI and bidirectional interaction

Addresses: Beyond current challenges toward proactive systems.

Current HAR operates unidirectionally: humans act, systems classify. Embodied AI demands bidirectional interaction: robots anticipate intentions, assistive devices predict needs 2–3 s ahead, and VR avatars respond naturally. Current forecasting reaches only 0.5–1s horizons [51], which is insufficient for proactive assistance.

Developing embodied motion intelligence requires predictive models that forecast intentions 3–5 s ahead via hierarchical motion plans, bidirectional synthesis that generates coordinated responses respecting safety constraints, and world models that integrate skeleton understanding with scene geometry and social norms. Achieving differential privacy ($\varepsilon < 1.0$) with less than 5% accuracy degradation remains challenging, as current privacy-preserving approaches incur 8–12% accuracy loss [3].

Success would enable assistive robots to achieve 80% proactive success in elderly care and VR avatars to improve social presence by 30–40%.

7. Conclusion

This survey has analyzed skeleton-based HAR through a generalization-centered lens, shifting focus from benchmark accuracy to deployment readiness. Our analysis produces four principal findings.

First, generalization challenges vary widely in severity and, in our sample, receive research attention that is roughly inversely related to their severity. Sensor noise induces the largest accuracy gap (46–60 percentage points between NTU and Kinetics), yet it is evaluated by the fewest methods. Conversely, the smallest gaps (cross-subject and cross-view) receive the most evaluation effort despite approaching benchmark saturation (Table 3). This pattern suggests the field optimizes for problems that are already largely solved while neglecting the hardest deployment barriers.

Second, every method family leaves at least three of six challenges unaddressed. The method–challenge suitability matrix (Table 4) confirms systematic coverage gaps: GCN-based methods dominate clean benchmarks but fail under noise; LLM integration enables open vocabulary recognition but violates real-time constraints; domain adaptation addresses environmental shifts but is evaluated in only 4–5 skeleton-specific papers. Real-world deployments face 3–4 challenges simultaneously, exposing a significant gap between single-challenge research and multi-challenge deployment.

Third, self-supervised learning and CNN-based representations emerge, on current evidence, as the most versatile paradigms. SSL addresses noise robustness (MaskCLR: 44.7% on Kinetics), cross-subject generalization (S-JEPA: 26.6% improvement under estimator shift), and label scarcity (93.9% with zero labels) simultaneously. CNNs combine noise robustness (PoseC3D: 47.7% on Kinetics, ahead of all reported GCN results there) with computational efficiency (4ms inference). Their combined pipeline (a CNN backbone with SSL pre-training) addresses four of the six challenges (C1, C2, C4, C6) and is the most deployment-ready approach currently available.

Fourth, evaluation infrastructure must evolve to match the challenges it aims to assess. Each current dataset tests at most two challenges simultaneously. No standard protocol mandates cross-dataset reporting, noise-robustness evaluation, or multidimensional scorecards. Constructing evaluation infrastructure commensurate with deployment requirements (cross-demographic benchmarks, robustness suites, multi-dimensional leaderboards) is a prerequisite for, not a consequence of, methodological progress.

Looking ahead, the convergence of skeleton-based HAR with physics-informed modeling, motion foundation models, and embodied AI presents opportunities to transform activity recognition from isolated classification into integrated motion intelligence. Realizing this potential requires the community to redirect efforts toward the largest generalization gaps, develop evaluation protocols that reflect deployment complexity, and build methods that address multiple challenges simultaneously rather than optimizing single benchmarks to diminishing returns.

CRediT authorship contribution statement

Yi Xia: Writing – original draft, Visualization, Investigation, Methodology, Conceptualization. **Sira Yongchareon:** Writing – review & editing, Supervision. **Raymond Lutui:** Writing – review & editing, Validation. **Quan Z. Sheng:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

No funding was received for conducting this study.

Data availability

No data were used for the research described in the article.

References

[1] M. Abdelfattah, A. Alahi, S-JEPA: a joint embedding predictive architecture for skeletal action recognition, in: European Conference on Computer Vision, Springer, 2024, pp. 367–384.
[2] M. Abdelfattah, M. Hassan, A. Alahi, Maskclr: attention-guided contrastive learning for robust action representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18678–18687.
[3] F.R. Albogamy, Federated learning for IOMT-enhanced human activity recognition with hybrid LSTM-GRU networks, Sensors 25 (2025) 907.
[4] L. Arrotta, G. Civitarese, C. Bettini, Dexar: deep explainable sensor-based activity recognition in smart-home environments, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6 (2022) 1–30.
[5] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, T. Delbruck, DHP19: dynamic vision sensor 3D human pose dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[6] J.W. Chang, M.H. Chen, H.S. Ma, H.L. Liu, Human movement science-informed multi-task spatio temporal graph convolutional networks for fitness action recognition and evaluation, Appl. Soft Comput. 164 (2024) 111963.
[7] C. Chen, R. Jafari, N. Kehtarnavaz, UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 168–172.
[8] D. Chen, M. Chen, P. Wu, M. Wu, T. Zhang, C. Li, Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition, Sci. Rep. 15 (2025) 4982.
[9] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, Y. Liu, Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities, ACM Computing Surveys (CSUR) 54 (2021) 1–40.
[10] Y. Chen, J. Guo, T. He, X. Lu, L. Wang, Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 778–786.
[11] Y. Chen, T. He, J. Fu, L. Wang, J. Guo, T. Hu, H. Cheng, Vision-language meets the skeleton: progressively distillation with cross-modal knowledge for 3D action representation learning, IEEE Trans. Multimed. 27 (2024) 2293–2303.
[12] Y. Chen, T. Jin, Z. Chen, Y. Lu, Q. Zhang, R. Wang, Skeleton-based action recognition with non-linear dependency modeling and Hilbert-schmidt independence criterion, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025.
[13] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.
[14] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.
[15] Q. Cheng, J. Cheng, Z. Ren, Q. Zhang, J. Liu, Multi-scale spatial–temporal convolutional neural network for skeleton-based action recognition, Pattern Analysis and Applications 26 (2023) 1303–1315.
[16] H.G. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, Infogcn: representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[17] M. Cormier, Y. Schmid, J. Beyerer, Enhancing skeleton-based action recognition in real-world scenarios through realistic data augmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 290–299.
[18] S. Cristina, V. Despotovic, R. Pérez-Rodríguez, S. Aleksic, Audio-and video-based human activity recognition systems in healthcare, IEEE Access 12 (2024) 8230–8245.
[19] R. Dai, S. Das, S. Sharma, L. Minciullo, L. Garattoni, F. Bremond, G. Francesca, Toyota smarthome untrimmed: real-world untrimmed videos for activity detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 2533–2550.
[20] M. Dalal, P. Mittal, A systematic review of deep learning-based models for elderly and human activity recognition, Comput. Sci. Rev. (2026), https://doi.org/10.1016/j.cosrev.2025.100715. In press.
[21] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi, H. Proença, Generative adversarial graph convolutional networks for human action synthesis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1150–1159.
[22] X. Ding, K. Yang, W. Chen, An attention-enhanced recurrent graph convolutional network for skeleton-based action recognition, in: Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, 2019, pp. 79–84.
[23] J. Do, M. Kim, Skateformer: skeletal-temporal transformer for human action recognition, in: European Conference on Computer Vision, Springer, 2024, pp. 401–420.

[24] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[25] H. Duan, M. Xu, B. Shuai, D. Modolo, Z. Tu, J. Tighe, A. Bergamo, Skeletr: towards skeleton-based action recognition in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13634–13644.
[26] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.
[27] C. Ellis, S.Z. Masood, M.F. Tappen, J.J. LaViola, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, Int. J. Comput. Vis. 101 (2013) 420–436.
[28] L. Franco, P. Mandica, B. Munjal, F. Galasso, Hyperbolic self-paced learning for self-supervised skeleton-based action representations, in: The Eleventh International Conference on Learning Representations, 2023, https://openreview.net/forum?id=3Bh6sRPKS3J.
[29] T. Guo, M. Liu, H. Liu, G. Wang, W. Li, Improving self-supervised action recognition from extremely augmented skeleton sequences, Pattern Recognit. 150 (2024) 110333.
[30] Z. Guo, Y. Liu, X. Wang, J. Chen, W. Zhang, MAR-GCN: meta-action refinement graph convolutional network for skeleton-based action recognition, Knowl.-Based Syst. 310 (2026) 112845.
[31] P. Gupta, D. Sharma, R.K. Sarvadevabhatla, Syntactically guided generative embeddings for zero-shot skeleton action recognition, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2021, pp. 439–443.
[32] Y. Hbali, S. Hbali, L. Ballihi, M. Sadgal, Skeleton-based human activity recognition for elderly monitoring systems, IET Comput. Vis. 12 (2018) 16–26.
[33] Y.H. Ho, J.H. Cheng, S.Y. Kuan, Z. Jiang, W. Chai, H.W. Huang, C.L. Lin, J.N. Hwang, RT-pose: a 4D radar tensor-based 3D human pose estimation and localization benchmark, in: European Conference on Computer Vision, Springer, 2024, pp. 107–125.
[34] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 28 (2018) 807–811.
[35] J. Hu, Y. Hou, Z. Guo, J. Gao, Global and local contrastive learning for self-supervised skeleton-based action recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (11) (2024) 10578–10589.
[36] R. Hu, X. Wang, X. Ding, Y. Zhang, X. Xin, W. Pang, S. Yu, Unsupervised domain adaptation for skeleton recognition with Fourier analysis, IEEE Internet Things J. 11 (24) (2024) 40166–40175.
[37] K.H. Huang, Y.B. Huang, Y.X. Lin, K.L. Hua, M. Tanveer, X. Lu, I. Razzak, GRA: graph representation alignment for semi-supervised action recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (2024) 11896–11905.
[38] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2013) 1325–1339.
[39] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, J. 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), IEEE, 2020, pp. 10990–10997.
[40] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.
[41] C. Ke, S. Liu, Y. Feng, S. Chen, Selective directed graph convolutional network for skeleton-based action recognition, Pattern Recognit. Lett. 190 (2025) 141–146.
[42] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3D action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3288–3297.
[43] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, T. Murakami, Mmact: a large-scale dataset for cross modal human action understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8658–8667.
[44] Y. Kong, Y. Fu, Human action recognition and prediction: a survey, Int. J. Comput. Vis. 130 (2022) 1366–1401.
[45] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Robot. Res. 32 (2013) 951–970.
[46] C. Li, P. Wang, S. Wang, Y. Hou, W. Li, Skeleton-based action recognition using LSTM and CNN, in: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2017, pp. 585–590.
[47] J. Li, T. Le, E. Shlizerman, Al-sar: active learning for skeleton-based action recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (11) (2023) 16966–16974.
[48] S.W. Li, Z.X. Wei, W.J. Chen, Y.H. Yu, C.Y. Yang, J.Y.J. Hsu, SA-dvae: improving zero-shot skeleton-based action recognition by disentangled variational autoencoders, in: European Conference on Computer Vision, Springer, 2024, pp. 447–462.
[49] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. 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, 2021, pp. 16266–16275.
[50] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 9–14.
[51] J. Lian, Y. Luo, X. Wang, L. Li, G. Guo, W. Ren, T. Zhang, Dual-STGAT: dual spatio-temporal graph attention networks with feature fusion for pedestrian crossing intention prediction, IEEE Trans. Intell. Transp. Syst. 26 (2025) 5396–5410.
[52] D. Liang, G. Fan, G. Lin, W. Chen, X. Pan, H. Zhu, Three-stream convolutional neural network with multi-task and ensemble learning for 3D action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[53] T. Liao, J.C. Chen, S.K. Jeng, C. Tai, Cross-domain knowledge transfer for skeleton-based action recognition based on graph convolutional gradient reversal layer, in: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2022, pp. 387–390.
[54] L. Lin, S. Song, W. Yang, J. Liu, MS2 l: multi-task self-supervised learning for skeleton based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2490–2498.
[55] L. Lin, L. Wu, J. Zhang, J. Liu, Idempotent unsupervised representation learning for skeleton-based action recognition, in: European Conference on Computer Vision, Springer, 2024, pp. 75–92.
[56] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, PKU-MMD: a large scale benchmark for skeleton-based human action understanding, in: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, Association for Computing Machinery, 2017, pp. 1–8, https://doi.org/10.1145/3132734.3132739
[57] C. Liu, Y. Jiang, C. Du, Z. Li, Enhancing action recognition from low-quality skeleton data via part-level knowledge distillation, Signal Process. 221 (2024) 109486.
[58] F. Liu, C. Wang, Z. Tian, S. Du, W. Zeng, Advancing skeleton-based human behavior recognition: multi-stream fusion spatiotemporal graph convolutional networks, Complex Intell. Syst. 11 (2025) 94.
[59] H. Liu, J. Tu, M. Liu, Two-stream 3D convolutional neural network for skeleton-based action recognition, arXiv preprint arXiv:1705.08106, 2017.
[60] J. Liu, C. Chen, M. Liu, Multi-modality co-learning for efficient skeleton-based action recognition, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4909–4918.
[61] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.Y. Duan, A.C. Kot, NTU rgb+ d 120: a large-scale benchmark for 3D human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2019) 2684–2701.
[62] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.
[63] J. Liu, G. Wang, P. Hu, L.Y. Duan, A.C. Kot, Global context-aware attention LSTM networks for 3D action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1656.
[64] J. Liu, B. Yin, J. Lin, J. Wen, Y. Li, M. Liu, Hdbn: a novel hybrid dual-branch network for robust skeleton-based action recognition, in: 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2024, pp. 1–6.
[65] M. Liu, H. Liu, Q. Hu, B. Ren, J. Yuan, J. Liu, J. Wen, 3D skeleton-based action recognition: a review, arXiv preprint arXiv:2506.00915, 2025.
[66] Q. Liu, Y. Wu, B. Li, Y. Ma, H. Li, Y. Yu, SHoTGCN: spatial high-order temporal GCN for skeleton-based action recognition, Neurocomputing (2025) 129697.
[67] R. Liu, Y. Chen, F. Gai, Y. Liu, Q. Miao, S. Wu, Local and global spatial-temporal transformer for skeleton-based action recognition, Neurocomputing (2025) 129820.
[68] R. Liu, Y. Liu, M. Wu, W. Xin, Q. Miao, X. Liu, L. Li, SG-CLR: semantic representation-guided contrastive learning for self-supervised skeleton-based action recognition, Pattern Recognit. (2025) 111377.
[69] Y. Liu, R. Liu, Y. Hu, M. Wu, W. Xin, Q. Miao, S. Wu, L. Li, A systematic review of skeleton-based action recognition: methods, challenges, and future directions, IEEE Trans. Neural Netw. Learn. Syst. (2025), https://doi.org/10.1109/TNNLS.2025.3632689
[70] M. Lu, H. Chen, Y. Zhang, L. Wang, J. Liu, Skeleton-prompt: skeleton-based action recognition with visual prompt tuning and 2D-to-3D pretraining, Pattern Recognit. 162 (2026) 111285.
[71] M. Lu, X. Lu, J. Liu, Momentum contrastive teacher for semi-supervised skeleton action recognition, IEEE Trans. Image Process. 34 (2025) 295–305.
[72] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, R. Stiefelhagen, drive&act: a multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2801–2810.
[73] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, A. Weber, MoCap database hdm05, Institut für Informatik II, Universität Bonn 2 (2007).
[74] M. Munaro, G. Ballin, S. Michieletto, E. Menegatti, 3D flow estimation for human action recognition from colored point clouds, Biol. Inspir. Cogn. Archit. 5 (2013) 42–51.
[75] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 122–132.
[76] N. Noor, F. Jametoni, J. Kim, H. Hong, I.K. Park, Efficient skeleton-based action recognition for real-time embedded systems, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5889–5897.
[77] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley mhad: a comprehensive multimodal human action database, in: 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2013, pp. 53–60.
[78] Q. Pan, Z. Zhao, X. Xie, J. Li, Y. Cao, G. Shi, View-normalized and subject-independent skeleton generation for action recognition, IEEE Trans. Circuits Syst. Video Technol. 33 (2022) 7398–7412.
[79] G. Paoletti, C. Beyan, A. Del Bue, SKELTER: unsupervised skeleton action denoising and recognition using transformers, Front. Comput. Sci. 5 (2023) 1203901.

[80] C. Plizzari, M. Cannici, M. Matteucci, Skeleton-based action recognition via spatial and temporal transformer networks, Comput. Vis. Image Underst. 208 (2021) 103219.
[81] W. Qian, Q. Huang, C. Li, Z. Chen, Y. Mao, EchoGCN: an echo graph convolutional network for skeleton-based action recognition, in: International Conference on Pattern Recognition, Springer, 2025, pp. 245–261.
[82] H. Qu, Y. Cai, J. Liu, LLMs are good action recognizers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18395–18406.
[83] H. Ramirez, S.A. Velastin, S. Cuellar, E. Fabregas, G. Farias, BERT for activity recognition using sequences of skeleton features and data augmentation with GAN, Sensors 23 (2023) 1400.
[84] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3D skeleton-based action recognition using learning method, Cyborg and Bionic Systems 5 (2024).
[85] A. Shahroudy, J. Liu, T.T. Ng, G. 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, 2016, pp. 1010–1019.
[86] J. Shen, J. Dudley, P.O. Kristensson, The imaginative generative adversarial network: automatic data augmentation for dynamic skeleton-based hand gesture and human action recognition, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–8.
[87] L. Shi, Y. Zhang, J. Cheng, H. 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, 2019, pp. 12026–12035.
[88] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Trans. Image Process. 29 (2020) 9532–9545.
[89] C. Si, W. Chen, W. Wang, L. Wang, T. 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, 2019, pp. 1227–1236.
[90] C. Silvano, D. Ielmini, F. Ferrandi, L. Fiorin, S. Curzel, L. Benini, F. Conti, A. Garofalo, C. Zambelli, E. Calore, S. Schifano, M. Palesi, G. Ascia, D. Patti, N. Petra, D. De Caro, L. Lavagno, T. Urso, V. Cardellini, G.C. Cardarilli, R. Birke, S. Perri, A survey on deep learning hardware accelerators for heterogeneous HPC platforms, ACM Comput. Surv. 57 (2025) 39.
[91] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).
[92] A. Sinha, D. Reilly, F. Bremond, P. Wang, S. Das, SKI models: skeleton induced vision-language embeddings for understanding activities of daily living, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 6931–6939.
[93] S. Song, C. Lan, J. Xing, W. Zeng, J. 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, 2017.
[94] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: a review, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 3200–3225.
[95] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBD images, in: 2012 IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 842–849.
[96] F.M. Thoker, J. Gall, Cross-modal knowledge distillation for action recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 6–10.
[97] H. Tian, J. Dickens, P. Payeur, Cross-graph domain adaptation for skeleton-based human action recognition, in: Proceedings of the Conference on Robots and Vision, PubPub, 2024.
[98] E. Todorov, T. Erez, Y. Tassa, MuJoCo: a physics engine for model-based control, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2012, pp. 5026–5033.
[99] Z. Tu, J. Zhang, H. Li, Y. Chen, J. Yuan, Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition, IEEE Trans. Multimed. 25 (2022) 1819–1831.
[100] I. Vernikos, E. Spyrou, Skeleton reconstruction using generative adversarial networks for human activity recognition under occlusion, Sensors 25 (2025) 1567.
[101] T.F.Y. Vicente, L. Hou, C.P. Yu, M. Hoai, D. Samaras, Large-scale training of shadow detectors with noisily-annotated shadow examples, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 816–832.
[102] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.
[103] J. Wang, X. Nie, Y. Xia, Y. Wu, S.C. Zhu, Cross-view action modeling, learning and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.
[104] L. Wang, D.Q. Huynh, P. Koniusz, A comparative review of recent kinect-based action recognition algorithms, IEEE Trans. Image Process. 29 (2019) 15–28.
[105] L. Wang, B. Sun, J. Robinson, T. Jing, Y. Fu, EV-action: electromyography-vision multi-modal action dataset, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 160–167.
[106] L. Wang, J. Zhang, W. Yang, S. Gu, S. Yang, 2D human skeleton action recognition with spatial constraints, IET Comput. Vis. 18 (2024) 968–981.
[107] P. Wang, W. Li, C. Li, Y. Hou, Action recognition based on joint trajectory maps with convolutional neural networks, Knowl.-Based Syst. 158 (2018) 43–53.
[108] X. Wang, Y. Dai, L. Gao, J. Song, Skeleton-based action recognition via adaptive cross-form learning, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1670–1678.
[109] X. Wang, L. Liu, F. Li, K. Lu, Active generation network of human skeleton for action recognition, Expert Syst. Appl. 281 (2025) 127529.
[110] Y. Wang, Y. Wu, W. He, X. Guo, F. Zhu, L. Bai, R. Zhao, J. Wu, T. He, W. Ouyang, Hulk: a universal knowledge translator for human-centric tasks, IEEE Trans. Pattern Anal. Mach. Intell. 47 (7) (2025) 5672–5689.
[111] K. Werling, J. Kaneda, T. Tan, R. Agarwal, S. Skov, T. Van Wouwe, S. Uhlrich, N. Bianco, C. Ong, A. Falisse, Addbiomechanics dataset: capturing the physics of human motion at scale, in: European Conference on Computer Vision, Springer, 2024, pp. 490–508.
[112] C. Wu, X.J. Wu, J. Kittler, T. Xu, S. Ahmed, M. Awais, Z. Feng, Scd-net: spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 5949–5957.
[113] L. Wu, C. Zhang, Y. Zou, SpatioTemporal focus for skeleton-based action recognition, Pattern Recognit. 136 (2023) 109231.
[114] W. Wu, C. Zheng, Z. Yang, C. Chen, S. Das, A. Lu, Frequency guidance matters: skeletal action recognition by frequency-aware mixed transformer, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4660–4669.
[115] Z. Wu, P. Sun, X. Chen, K. Tang, T. Xu, L. Zou, X. Wang, M. Tan, F. Cheng, T. Weise, SelfGCN: graph convolution network with self-attention for skeleton-based action recognition, IEEE Trans. Image Process. 33 (2024) 4391–4403.
[116] L. Xia, C.C. Chen, J.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, IEEE, 2012, pp. 20–27.
[117] W. Xiang, C. Li, Y. Zhou, B. Wang, L. Zhang, Generative action description prompts for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10276–10285.
[118] J. Xie, Y. Meng, Y. Zhao, A. Nguyen, X. Yang, Y. Zheng, Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 6225–6233.
[119] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, Q. Miao, Transformer for skeleton-based action recognition: a review of recent advances, Neurocomputing 537 (2023) 164–186.
[120] H. Xu, Y. Gao, Z. Hui, J. Li, X. Gao, Language knowledge-assisted representation learning for skeleton-based action recognition, IEEE Trans. Multimed. 27 (2025) 5784–5799.
[121] Q. Yan, Y. Hu, A transformer-based unsupervised domain adaptation method for skeleton behavior recognition, IEEE Access 11 (2023) 51689–51700.
[122] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[123] T. Yan, W. Zeng, Y. Xiao, X. Tong, B. Tan, Z. Fang, Z. Cao, J.T. Zhou, Crossglg: LLM guides one-shot skeleton-based 3D action recognition in a cross-level manner, in: European Conference on Computer Vision, Springer, 2024, pp. 113–131.
[124] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, F. Brémond, View-invariant skeleton action representation learning via motion retargeting, Int. J. Comput. Vis. 132 (2024) 2351–2366.
[125] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C.X. Lu, L. Xie, Mm-fi: multi-modal non-intrusive 4D human dataset for versatile wireless sensing, Adv. Neural Inf. Process. Syst. 36 (2023) 18756–18768.
[126] S. Yang, X. Wang, L. Gao, J. Song, Mke-gcn: multi-modal knowledge embedded graph convolutional network for skeleton-based action recognition in the wild, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 01–06.
[127] L. Yuan, Z. He, Q. Wang, L. Xu, X. 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, IEEE, 2022, pp. 1–6.
[128] R. Yue, Z. Tian, S. Du, Action recognition based on RGB and skeleton data sets: a survey, Neurocomputing 512 (2022) 287–306.
[129] X. Yun, C. Xu, K. Riou, K. Dong, Y. Sun, S. Li, K. Subrin, P. Le Callet, Behavioral recognition of skeletal data based on targeted dual fusion strategy, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 6917–6925.
[130] J. Zhang, L. Lin, S. Yang, J. Liu, Self-supervised skeleton-based action representation learning: a benchmark and beyond, Int. J. Comput. Vis. (2026), https://doi.org/10.1007/s11263-025-02644-8
[131] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. 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, 2017, pp. 2117–2126.
[132] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, N. Zheng, EleAtt-RNN: adding attentiveness to neurons in recurrent neural networks, IEEE Trans. Image Process. 29 (2019) 1061–1073.
[133] Y. Zhang, Z. Zhao, W. Li, L. Duan, Multi-scale enhanced active learning for skeleton-based action recognition, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[134] B. Zhao, S. Li, Y. Gao, IndRNN based long-term temporal recognition in the spatial and frequency domain, in: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, 2020, pp. 368–372.

[135] Y. Zhou, Z.Q. Cheng, C. Li, Y. Fang, Y. Geng, X. Xie, M. Keuper, Hypergraph transformer for skeleton-based action recognition, arXiv preprint arXiv:2211.09590, 2022.
[136] Y. Zhou, H. Duan, A. Rao, B. Su, J. Wang, Self-supervised action representation learning from partial spatio-temporal skeleton sequences, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 3825–3833.
[137] Y. Zhou, X. Yan, Z.Q. Cheng, Y. Yan, Q. Dai, X.S. Hua, Blockgcn: redefine topology awareness for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2049–2058.
[138] A. Zhu, Q. Ke, M. Gong, J. Bailey, Part-aware unified representation of language and skeleton for zero-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18761–18770.
[139] Y. Zhu, H. Shuai, G. Liu, Q. Liu, Multilevel spatial–temporal excited graph network for skeleton-based action recognition, IEEE Trans. Image Process. 32 (2022) 496–508.
[140] D. Zhuang, M. Jiang, J. Kong, Time-to-space progressive network using overlap skeleton contexts for action recognition, Signal Process. 207 (2023) 108953.
[141] T. Zhuang, Z. Qin, Y. Ding, Z. Qin, J. Geng, Y. Liu, K.K.R. Choo, DSDC-GCN: decoupled static-dynamic co-occurrence graph convolutional networks for skeleton-based action recognition, IEEE Trans. Circuits Syst. Video Technol. 35 (3) (2024) 2101–2117.