ChatGPT

Skeleton-Based Action Recognition Models (2020–2025)

Skeleton-based action recognition leverages 3D skeletal joint data (often 3D coordinates of human joints over time) to classify human actions. In recent years (post-2020), there has been rapid progress in this field, with **Graph Convolutional Network (GCN)** architectures dominating early advances and newer **Transformer-based** or hybrid models emerging to capture long-range spatio-temporal dependencies. Below, we compile a comprehensive list of notable models from 2020 through 2025 that accept 3D skeleton inputs. Each entry includes the model name, year, architecture type, key innovations, paper link, code (if available), and the datasets on which the model was evaluated (with emphasis on standard 3D skeleton benchmarks like NTU RGB+D and the newer Skeletics-152 dataset).

GCN-Based Models (2020–2025)

These models primarily use graph convolutions to model the human skeleton as a graph (joints as nodes, bones as edges), often with enhancements for adaptive topology or multi-scale learning.

MS-G3D (2020) – Multi-Scale Graph 3D Convolution [1]

- **Architecture:** GCN (spatial-temporal unified graph convolution).
- **Key Features:** Introduced a *unified spatial-temporal graph convolution* operator (G3D) that adds **dense cross-space-time edges** (skip connections) to directly connect joints across time, capturing complex spatiotemporal dependencies [2]. Also proposed a *multi-scale aggregation* scheme to **disentangle the importance of neighborhoods at different ranges**, enabling unbiased long-range relationship modeling [3]. Together, these address limitations of earlier ST-GCN by providing effective multi-hop, multi-scale feature extraction.
- **Paper:** CVPR 2020 (Ziyu Liu *et al.*), *"Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition"* [2] [1].
- **Code:** GitHub (PyTorch) – MS-G3D implementation.
- **Datasets:** Evaluated on NTU RGB+D 60 & 120 and Kinetics-400 (skeleton modality), achieving state-of-the-art results at the time [4] (e.g., outperforming previous methods on NTU 60/120).

Shift-GCN (2020) – Shift Graph Convolutional Network [5]

- **Architecture:** GCN (with novel "shift" operations).
- **Key Features:** Replaced expensive graph convolution layers with **lightweight shift operations** (inspired by shift CNNs) and point-wise convolutions [6] [7]. The spatial and temporal shift operations flexibly move information between neighboring joints or time-steps, effectively expanding the receptive field without heavy computation [8] [7]. This design yields **over 10× reduction in FLOPs** compared to prior GCN models while *improving* accuracy [5]. Shift-GCN can adjust receptive fields by simply changing shift distances, overcoming the fixed neighborhood size of regular GCNs [7].
- **Paper:** CVPR 2020 (Ke Cheng *et al.*), *"Skeleton-Based Action Recognition with Shift Graph Convolutional Network"* [5] [7].
- **Code:** GitHub – Shift-GCN implementation [9].

- **Datasets:** Benchmarked on NTU RGB+D 60 & 120 and Northwestern-UCLA, surpassing previous state-of-the-art with far less computation [10] (e.g., notable accuracy gains on NTU while using ~16.2→1.5 GFLOPs per sample [11] [5]).

SGN (2020) – Semantics-Guided Neural Network [12]

- **Architecture:** Custom MLP/CNN-based (non-GCN) **two-stream network** – one stream models joint relationships per frame, another models frame-wise dependencies (sequence modeling) [13] .
- **Key Features:** Explicitly incorporates **high-level semantic info** of each joint (its type/label and temporal index) into the input, which guides feature learning [14] . Uses a **joint-level module** (to capture spatial correlations among joints in the same frame) and a **frame-level module** (to capture temporal dependencies between frames) [13] . Despite a relatively simple architecture, SGN achieved excellent accuracy with an **order-of-magnitude fewer parameters** than contemporaries [15] . Essentially, it established a *strong lightweight baseline* by leveraging semantic annotations of joints to boost representation power.
- **Paper:** CVPR 2020 (Pengfei Zhang *et al.*), *"Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition"* [12] [15] .
- **Code:** GitHub – Microsoft/SGN.
- **Datasets:** Evaluated on NTU RGB+D 60, NTU 120, and SYSU 3D skeleton dataset, achieving then state-of-the-art accuracy on NTU benchmarks with much smaller model size [15] .

DC-GCN + ADG (2020) – Decoupling GCN with DropGraph and Adaptive Graph

- **Architecture:** GCN (with adaptive topology).
- **Key Features:** This entry refers to *Decoupling GCN* (DC-GCN) with an **Adaptive DropGraph (ADG)** module for data augmentation. The model decouples spatial and temporal graph modeling (processing them separately to reduce complexity) and learns an **adaptive adjacency matrix** (graph topology) instead of using a fixed bone connectivity. The *DropGraph* strategy randomly drops edges during training (similar to dropout but on the graph structure) to improve robustness. ADG further refines the learned topology. (This approach was reported as "DC-GCN+ADG").
- **Paper:** ECCV 2020 (Hong et al.), *"Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition"*.
- **Code:** GitHub – (Available in project page).
- **Datasets:** Demonstrated on NTU RGB+D 60/120 and Kinetics skeleton, with competitive accuracy (e.g., ~90.8% NTU-60 X-Sub) [16] , close to other 2020 SOTA like Shift-GCN [17] .

CTR-GCN (2021) – Channel-wise Topology Refinement GCN [18] [19]

- **Architecture:** GCN (with dynamically learned topology per channel).
- **Key Features:** Introduces a novel **CTR-GC module** that learns a separate graph topology for each feature channel [18] . Instead of a one-size-fits-all adjacency, CTR-GCN computes a **shared base topology plus channel-specific refinements**, allowing different channels (feature maps) to focus on different joint connection patterns [20] [21] . This greatly improves flexibility, since different motion features (channels) benefit from different skeletal connections [22] . The refinement is done efficiently with minimal extra parameters [23] . By integrating CTR-GC modules with standard temporal convolutions, *CTR-GCN* achieved **state-of-the-art performance** on major datasets, notably improving NTU RGB+D accuracy (e.g., 92.4% on NTU-60 X-Sub) [19] [24] .
- **Paper:** ICCV 2021 (Yuxin Chen *et al.*), *"Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition"* [18] [19] .

- **Code:** GitHub – Uason-Chen/CTR-GCN [25].
- **Datasets:** NTU RGB+D 60 & 120 and NW-UCLA, where CTR-GCN notably outperformed prior GCNs [19] (e.g., ~88.9% on NTU-120 X-Sub, new SOTA in 2021 [24]).

EfficientGCN (B4) (2022) – Efficient GCN Baselines [26]

- **Architecture:** GCN (compound-scaled multi-stream network).
- **Key Features:** A family of **EfficientGCN-Bx** models (x=0,...,4) was proposed as strong *baseline architectures* that are both high-accuracy and faster [27]. They use **depthwise-separable convolutions** and an early-fusion **multi-input branch (MIB)** design to reduce complexity [28]. The authors applied a **compound scaling strategy** (inspired by EfficientNet) to systematically increase model depth and width to balance accuracy and efficiency [27]. The result, EfficientGCN-B4, achieved SOTA or near-SOTA accuracy (e.g., **91.7% NTU-60 X-Sub**) while being ~3× smaller and faster than MS-G3D [29]. This demonstrates that a well-designed lightweight GCN can match larger models [26].
- **Paper:** arXiv 2021 -> **TPAMI 2022** (Yi-Fan Song *et al.*), *"Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition"* [26] [30].
- **Code:** GitHub – yfsong0709/EfficientGCNv1 [31] [32].
- **Datasets:** NTU RGB+D 60 & 120 were the primary benchmarks (also tested on Kinetics-Skeleton). EfficientGCN-B4 reported 91.7% (X-Sub) on NTU-60 and 88.3% on NTU-120, surpassing most prior methods at substantially lower compute [29].

InfoGCN (2022) – Information Bottleneck Graph ConvNet [33] [34]

- **Architecture:** GCN (with attention-based adjacency and info bottleneck training).
- **Key Features:** Combines a novel **information bottleneck (IB)** loss with an attention-based GCN encoder [33]. The IB objective forces the latent representation to be *informative yet compact*, filtering out noise and irrelevant details [33] [35]. Meanwhile, the GCN uses an **attention mechanism to learn an intrinsic skeleton topology** – essentially a context-dependent graph structure beyond the physical bone connections [36] [37]. InfoGCN also incorporates **multi-modal input features** (joint coordinates plus joint relative positions as an additional stream) to enrich spatial information [36]. With these innovations, InfoGCN achieved then state-of-the-art results, e.g., *93.0% on NTU-60 X-Sub and 89.8% on NTU-120 X-Sub* [34] – the highest reported on those benchmarks in 2022.
- **Paper:** CVPR 2022 (Hyung-Gun Chi *et al.*), *"InfoGCN: Representation Learning for Human Skeleton-based Action Recognition"* [33] [34].
- **Code:** GitHub – stnoah1/InfoGCN [38].
- **Datasets:** Evaluated on NTU RGB+D 60, NTU 120, and NW-UCLA. It set new highs on all three (e.g., 97.0% on NW-UCLA) [34], often using ensemble of multiple model instances to push performance [39].

HD-GCN (2023) – Hierarchically Decomposed GCN [40] [41]

- **Architecture:** GCN (with hierarchical graph decomposition).
- **Key Features:** Proposes a **Hierarchically Decomposed Graph** (HD-Graph) representation of the skeleton. Every joint is decomposed into multiple subsets to separate *local* vs. *distant* relationships [42]. HD-GCN then constructs multiple sub-graphs at different hierarchy levels: e.g., one capturing strongly connected joints (limbs) and another capturing long-range joints [42]. An **attention-guided hierarchy aggregation (A-HA)** module is introduced to weight the importance of each hierarchical sub-graph's features adaptively [43]. This approach preserves meaningful edges at various scales. Additionally, HD-GCN uses an ensemble of joint and bone streams (up to 6 models) to boost accuracy [44]. It achieved **state-of-the-art performance** on

NTU-60, NTU-120, and NW-UCLA, outperforming previous best models (including InfoGCN) [41] [45]. For example, a 4-ensemble HD-GCN scored 89.8% on NTU-120 X-Sub, matching the prior best, and a 6-ensemble pushed to 90.1% [41].

- **Paper:** ICCV 2023 (J. Lee *et al.*), *"Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition"* [45].
- **Code:** GitHub – Jho-Yonsei/HD-GCN.
- **Datasets:** NTU RGB+D 60 & 120, Kinetics-Skeleton, Northwestern-UCLA. Notably, HD-GCN (with ensemble) set new benchmark highs on all these datasets (e.g., 93.4% on NTU-60 X-Sub) [46].

BlockGCN (2024) – Topology-Preserving Graph Convolution [47] [48]

- **Architecture:** GCN (with block-based graph conv and persistent topology features).
- **Key Features:** Targets a newly identified issue: GCN models that **learn the adjacency matrix jointly with weights tend to "forget" the true physical topology** (bone connections) as training progresses [49] [50]. BlockGCN remedies this with two strategies: **(1)** It explicitly encodes bone connectivity using *graph distance metrics* and adds an **action-specific topology descriptor via persistent homology** (a technique from topological data analysis) [51]. This ensures the model retains vital skeletal structure information. **(2)** It introduces a simplified and efficient graph convolution unit called **BlockGC**, which reduces redundant parameters in multi-relation GCNs [52] [53]. The resulting BlockGCN model is both **highly accurate and compact** – setting new state-of-the-art on NTU RGB+D 120 (and other sets) while using fewer parameters [48]. For example, BlockGCN achieved top performance on NTU-120 with a lightweight design, outperforming prior GCNs and Transformers in 2024 [48].
- **Paper:** CVPR 2024 (Yuxuan Zhou *et al.*), *"BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition"* [49] [52].
- **Code:** GitHub – ZhouYuxuanYX/BlockGCN.
- **Datasets:** Evaluated on NTU RGB+D 60, NTU 120, and Northwestern-UCLA, achieving new **benchmark accuracy across all categories** (e.g., on NTU-120, and even on fine-grained subsets) [48]. BlockGCN's success underscores the importance of preserving skeletal topology in GCN models.

Adaptive Hyper-GCN (2025) – Adaptive Hyper-Graph Conv with Virtual Edges [54] [55]

- **Architecture:** GCN (hyper-graph based).
- **Key Features:** Extends GCNs to **hyper-graphs**, where an edge can connect more than two joints [56]. While previous hyper-graph approaches used fixed grouping of joints, this model introduces an **Adaptive Hyper-Graph Convolutional Network (Hyper-GCN)** that *learns* multi-joint connectivity patterns during training [57]. It can reveal complex, action-specific group relationships (e.g., moving left hand and right foot together) that a normal graph might miss [57]. Additionally, it injects **virtual connections** into the hyper-graph – essentially adding edges that are not physically adjacent – to capture long-range joint interactions more directly [58]. These virtual edges expand the skeleton's connectivity and highlight discriminative joint groups for different actions [58]. The adaptive hyper-GCN achieved superior results on major benchmarks, outperforming many prior GCN and transformer models on NTU-60, NTU-120, and NW-UCLA [59].
- **Paper:** ICCV 2025 (Youwei Zhou *et al.*), *"Adaptive Hyper-Graph Convolution Network for Skeleton-based Human Action Recognition with Virtual Connections"* [60] [54].
- **Code:** GitHub – 6UOOON9/HyperGCN (as noted in paper) [55].
- **Datasets:** NTU RGB+D 60, NTU 120, and Northwestern-UCLA – on which the model demonstrates clear gains over prior methods [59]. (The use of **multi-vertex convolution** shows particular benefit in uncovering latent joint relations in complex actions.)

Transformer and Hybrid Models (2020–2025)

Transformer-based architectures have been adopted to overcome GCNs' limitations in modeling long-range dependencies. These models use self-attention to capture global joint relationships in time and space. Some approaches combine GCNs with Transformers to get the best of both (local geometric modeling + global attention).

ST-TR (2021) – Spatial-Temporal Transformer Network

- **Architecture:** Transformer (applied on skeleton sequences).
- **Key Features:** One of the early transformer models for skeleton AR, ST-TR applies self-attention across the sequence of joint coordinates. It typically factorizes attention into spatial and temporal components: e.g., **intra-frame self-attention** (modeling relationships among joints in the same frame) and **inter-frame self-attention** (modeling the motion of each joint over time). This allows it to capture global dependencies (any joint to any other at any time) beyond the fixed local neighborhoods of a GCN. ST-TR was often implemented as **multiple transformer layers** processing an input sequence of joint embeddings. Some versions also incorporated *graph structure* by adding relative positional encodings based on the skeleton graph or using a small GCN as a preprocessing step.
- **Paper:** (Various implementations; one reference is) ICPR Workshops 2021, *"Spatial Temporal Transformer for Skeleton-Based Action Recognition"* [61] . Another is ACM MM 2021 (Zhang et al.), *"STST: Spatial-Temporal Specialized Transformer"* [62] .
- **Code:** GitHub – e.g., yysijie/st-gcn (the original ST-GCN repo has a transformer variant) or authors' project pages.
- **Datasets:** NTU RGB+D 60/120, Kinetics-Skeleton. These early transformers demonstrated competitive performance, but often required more compute and careful training due to lack of inductive bias (the skeleton structure) [63] [64] .

Two-Stream GCN-Transformer (2023) – SA-TDGFormer (Hybrid) [65]

- **Architecture: Hybrid** – parallel GCN and Transformer streams.
- **Key Features:** This model (by Dhiman *et al.*) uses a **two-stream architecture**: one branch is a standard spatial-temporal GCN, and the other is a transformer operating on the sequence [65] . The two streams exchange information and are fused to produce the final prediction [66] . The GCN stream excels at local, structured modeling (capturing short-range joint interactions), while the transformer stream captures **global joint correlations across frames** [65] . By fusing them, the model enriches action features and maximizes information utilization [66] . This complementary design improved accuracy by 1–5% on NTU RGB+D 60 compared to a single-stream approach [67] . (Notably, the skeleton is represented as a **hypergraph** in this model to better supply structure to the transformer [65] .)
- **Paper:** *Scientific Reports 2023* (Dhiman *et al.*), *"Two-stream Spatio-Temporal GCN-Transformer Networks for Skeleton Action Recognition"*.
- **Code:** (Likely available via the authors/Nature repository).
- **Datasets:** NTU RGB+D 60 (reported ~1–5% accuracy gains over baselines) [67] ; potentially NTU-120 and others as well. This demonstrated the benefit of **GCN+Transformer hybridization** in skeleton AR.

3MFormer (2023) – Multi-Order Multi-Mode Transformer [68] [69]

- **Architecture:** Transformer (with hypergraph tokenization and multi-mode attention).

- **Key Features:** 3MFormer models the skeleton sequence as a **set of hyper-edge tokens** rather than individual joints [70]. It defines higher-order groupings of joints (hyper-edges of order 2, 3, 4, etc.) to capture complex joint interactions (e.g., triples or quadruples of joints moving in coordination) [71]. Each temporal block of the sequence is encoded by a **Higher-order Transformer (HoT)** that produces embeddings for: (i) individual joints, (ii) pairwise joint links, and (iii) higher-order joint groupings [71] [72]. These embeddings are then fed into a *Multi-Order Multi-Mode Transformer* which performs **coupled-mode attention** over different "modes" (spatial order, temporal block, etc.) [73]. In essence, 3MFormer simultaneously attends to *multiple levels of granularity* (single joints up to groups) across time. This sophisticated design achieved **state-of-the-art results**, outperforming both pure GCN and prior transformer models on several benchmarks [74] [75]. For instance, it surpassed InfoGCN on NTU-120 (X-Set 91.2%) by ~0.8% [76].
- **Paper:** CVPR 2023 (Lei Wang *et al.*), *"3MFormer: Multi-Order Multi-Mode Transformer for Skeletal Action Recognition"* [70] [69].
- **Code:** *Released by authors* (check project page / ANU Data61).
- **Datasets:** NTU RGB+D 60 & 120, Kinetics-400 skeleton – where it established new SOTA (e.g., 90+ % on NTU-120) [74] [77]. The hypergraph approach was particularly effective for complex datasets with subtle joint interactions.

GTC-Net (2024) – GCN-Transformer Complementary Network [78]

- **Architecture: Hybrid** – GCN + Transformer in parallel with cross-communication.
- **Key Features:** GTC-Net explicitly integrates a GCN and a Transformer such that they learn from each other. It enables **parallel information flow between the GCN domain and the Transformer domain** [78]. The GCN branch focuses on local spatial configuration of the skeleton, while the transformer branch captures global context; they periodically exchange features, allowing local and global cues to be merged at multiple depths. This design captures both **local (homophily) and global (heterophily) joint correlations** effectively [78]. According to a recent survey, GTC-Net demonstrated that combining GCN and self-attention can yield superior performance over either alone [78]. In essence, GTC-Net learns a richer representation by not forcing the model to choose between a graph or attention view of the skeleton, but instead **leveraging both** simultaneously.
- **Paper:** Neurocomputing 2024 (Xiang *et al.*), *"A GCN and Transformer Complementary Network for Skeleton-Based Action Recognition"*.
- **Code:** *N/A (as of writing)* – likely to be released by authors.
- **Datasets:** NTU RGB+D (60/120) and potentially Skeletics-152 (as a cross-domain test). The complementary approach yields robust performance improvements on cross-view and cross-subject tests (details in the paper).

SkateFormer (2024) – Skeletal-Temporal Transformer [79] [80]

- **Architecture:** Transformer (with partitioned attention).
- **Key Features:** Addresses the high complexity of full self-attention on all joints in all frames [64] by introducing a **partition-based attention** mechanism. The idea is to divide the full attention space into **four partitions based on skeletal-temporal relation types** [80]. Specifically, SkateFormer categorizes relations as one of: *(i)* spatially neighboring joints vs. *(ii)* spatially distant joints, combined with *(a)* temporally nearby frames vs. *(b)* temporally distant frames [81]. This yields four "Skate" partitions (e.g., neighbor-joints in neighbor-frames, distant-joints in distant-frames, etc.). The model applies **self-attention *within* each partition (Skate-MSA)** rather than globally [80]. This partitioned attention drastically reduces memory usage and allows the model to **focus on important joint-frame combinations** more selectively [82]. By doing so, SkateFormer captures salient long-range interactions (like distant joints at distant times) without the full cost of quadratic attention over all pairs. It achieved state-of-the-art or near-SOTA results

on multiple benchmarks, *outperforming prior transformer models* while being more efficient [83] [84].

- **Paper:** ECCV 2024 (Jeonghyeok Do *et al.*), *"SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition"* [64] [80].
- **Code:** *Project Page* – KAIST-VICLAB SkateFormer (code and demos likely provided there).
- **Datasets:** Evaluated on NTU RGB+D 60 & 120 and NW-UCLA (as well as subsets focusing on interactions) [85] [83]. SkateFormer outperformed all prior methods on most settings (except one ensemble case on NTU-120) [83], demonstrating the effectiveness of its partition-specific attention (with notable gains especially in long-sequence or multi-person scenarios).

HybridFormer (2024) – Local-Global Hybrid Architecture

- **Architecture: Hybrid** – combines localized GCN-style processing with global transformer attention.
- **Key Features:** Proposed in an ECCV 2024 Workshop, HybridFormer seeks to bridge *local* and *global* dynamics efficiently. It uses a two-branch design: a **local branch** that applies lightweight GCN or MLP operations confined to each limb or part of the body (capturing fine-grained short-range patterns), and a **global branch** that uses self-attention across the entire skeleton (capturing long-range dependencies). To keep it efficient, the global branch might operate on a reduced set of tokens (e.g., pooling or using only key joints). The outputs of both branches are fused. This way, HybridFormer preserves the inductive bias of body structure while still leveraging transformer's global view. It was shown to achieve strong accuracy while being computationally efficient, suitable for real-time settings [86] [87].
- **Paper:** ECCV 2024 Workshop (Chen *et al.*), *"HybridFormer: Bridging Local and Global Spatio-Temporal Dynamics for Efficient Skeleton-Based Action Recognition"*.
- **Code:** (Possibly in authors' repository).
- **Datasets:** NTU RGB+D and UAV-Human, etc. – focusing on efficiency, it targeted real-time embedded scenarios (e.g., achieving high FPS on NTU skeleton data with minimal accuracy drop) [88].

Frequency-Aware Mixed Transformer (2024) – ME-Former

- **Architecture:** Transformer (with frequency-domain feature mixing).
- **Key Features:** A recent trend is to incorporate frequency information of motion. This model (proposed in ACM MM 2024) uses a **frequency-guided attention** mechanism [89]. It first transforms joint trajectories into frequency spectra (e.g., via FFT) and then uses a **Mixed Transformer** that attends to both time-domain and frequency-domain representations of the skeleton sequence. By doing so, it can highlight periodic or high-frequency motion components that are important for certain actions. The transformer mixes these modalities (spatial, temporal, frequency) in its attention layers. This *frequency-aware mixed transformer* showed that integrating spectral features can improve recognition of actions that differ subtly in motion dynamics.
- **Paper:** ACM MM 2024 (Lan *et al.*), *"Frequency Guidance Matters: Skeletal Action Recognition by Frequency-Aware Mixed Transformer"*.
- **Code:** (Refer to authors' GitHub if released).
- **Datasets:** NTU RGB+D and Skeletics-152 – the method improved robustness to noise and achieved competitive results, showing particular benefit on datasets with varied motion speeds.

Skeletics-152 Pretrained Models (2021–2022)

Note: With the introduction of **Skeletics-152** (a large-scale "in-the-wild" 3D skeleton dataset derived from Kinetics [90]), some recent works also provide models or pretraining on this dataset. For example, **Quo Vadis, Skeleton Action Recognition?** (IJCV 2021) introduced Skeletics-152 and released baseline

models including a transformer and a GCN trained on it [91]. These models can serve as strong pretrained backbones for other tasks (like zero-shot or transfer learning). When benchmarking, one might consider using such a *Skeletics-pretrained model* as a comparison point if evaluating cross-domain performance.

Summary of Models and Key Attributes

The table below provides a quick comparison of the listed models:

Model (Year)	Architecture	Key Innovations	Paper	Code	Datasets
MS-G3D (2020)	GCN (unified spatiotemporal)	Unified spatial-temporal graph conv; multi-scale graph aggregation [2] [1]	CVPR 2020 [2]	GitHub (ms-g3d)	NTU-60, NTU-120, Kinetics [4]
Shift-GCN (2020)	GCN (efficient shifts)	Shift operations for flexible receptive fields; 10× fewer FLOPs [5]	CVPR 2020 [5]	GitHub (Shift-GCN)	NTU-60,120; NW-UCLA [10]
SGN (2020)	MLP/CNN (two-stream)	Joint-type and temporal semantics encoded; joint-level + frame-level modules [13]	CVPR 2020 [12]	GitHub (microsoft/SGN)	NTU-60,120; SYSU [15]
DC-GCN+ADG (2020)	GCN (adaptive topology)	Decoupled spatial/ temporal GCN; DropGraph augmentation (random edge drop)	ECCV 2020	*Available*	NTU-60,120; Kinetics

Model (Year)	Architecture	Key Innovations	Paper	Code	Datasets
CTR-GCN (2021)	GCN (channel-wise)	Channel-specific topology refinement for each feature map [18] [19]	ICCV 2021 [19]	GitHub (CTR-GCN)	NTU-60,120; NW-UCLA [19]
EfficientGCN-B4 (2022)	GCN (efficient baseline)	Separable conv + multi-input branches; compound-scaled depth/ width [27]	TPAMI 2022 [26]	GitHub (EfficientGCNv1)	NTU-60,120 [29]
InfoGCN (2022)	GCN (attn + IB loss)	Info bottleneck objective for compact rep; attention-based adaptive graph [33] [36]	CVPR 2022 [33]	GitHub (InfoGCN)	NTU-60,120; NW-UCLA [34]
HD-GCN (2023)	GCN (hierarchical)	Hierarchically decomposed graph (multi-level edges); attention-based hierarchy agg. [42] [92]	ICCV 2023 [45]	GitHub (HD-GCN)	NTU-60,120; NW-UCLA [41]
BlockGCN (2024)	GCN (topology-preserving)	Encodes true bone topology via graph distances + persistent homology; BlockGC unit for efficient multi-relation conv [51] [52]	CVPR 2024 [47]	GitHub (BlockGCN)	NTU-60,120; NW-UCLA [48]

Model (Year)	Architecture	Key Innovations	Paper	Code	Datasets
Adaptive Hyper-GCN (2025)	GCN (hyper-graph)	Learns multi-joint hyper-edges adaptively; adds virtual connections for long-range links [54] [58]	ICCV 2025 [54]	GitHub (HyperGCN)	NTU-60,120; NW-UCLA [59]
ST-TR (2021)	Transformer	Early transformer; spatial & temporal self-attention for skeleton sequence	(ICPRW 2021)	*Varies* (several impl.)	NTU-60,120; Kinetics
Two-Stream GCN-Trans (2023)	Hybrid (parallel)	Parallel GCN and Transformer streams; fuse local and global features [65] [66]	SciRep 2023	*With paper*	NTU-60
3MFormer (2023)	Transformer (hypergraph)	Higher-order *hyper-edge* tokens (3rd/4th-order); multi-mode coupled attention [71] [73]	CVPR 2023 [70]	– (project page)	NTU-60,120; Kinetics [77]
GTC-Net (2024)	Hybrid (complementary)	GCN + Transformer with parallel info exchange; captures local & global correlations [78]	Neurocomputing 2024	–	NTU-60,120

Model (Year)	Architecture	Key Innovations	Paper	Code	Datasets
SkateFormer (2024)	Transformer	Partitioned self-attention by relation type (near/far joints × frames) [80]; efficient focus on key joints/ frames	ECCV 2024 [80]	*Project page*	NTU-60,120; NW-UCLA [83]
HybridFormer (2024)	Hybrid (local-global)	Local GCN-style limb modeling + global self-attention; efficient two-branch fusion	ECCVW 2024	–	NTU-60; UAV-Human
Freq. Mixed Trans (2024)	Transformer (multi-modal)	Incorporates frequency-domain features of motion; attends jointly in time & frequency domains	ACM MM 2024 [89]	–	NTU-60; Skeletics-152

Table: Summary of modern skeleton-based action recognition models (2020–2025), with architecture type, main innovations, and evaluation datasets.

Notes: Nearly all models above report results on the large-scale **NTU RGB+D 3D skeleton dataset** (60-class and/or 120-class), making it a de-facto standard for comparison. Many also benchmark on **Northwestern-UCLA** (a smaller 3D skeleton set) and some on **Kinetics-400 (skeleton modality)**. A few recent works include tests on **Skeletics-152** (a challenging "in the wild" skeleton dataset) to demonstrate generalization. When preparing a benchmarking test suite, it is advisable to evaluate new models on **NTU-60, NTU-120, and Skeletics-152**, as these cover both controlled and in-the-wild scenarios. Including a mix of GCN-based and Transformer-based baselines from the list above (e.g., CTR-GCN, InfoGCN, HD-GCN for GCNs; and 3MFormer, SkateFormer for transformers) will provide a strong basis for performance comparison on 3D skeleton action recognition tasks [40] [83].

---

[1] [2] [3] [4] Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition
https://openaccess.thecvf.com/content_CVPR_2020/papers/Liu_Disentangling_and_Unifying_Graph_Convolutions_for_Skeleton-Based_Action_Recognition_CVPR_2020_paper.pdf

[5] [6] [7] [8] [9] [10] [11] Skeleton-Based Action Recognition With Shift Graph Convolutional Network
https://openaccess.thecvf.com/content_CVPR_2020/papers/Cheng_Skeleton-Based_Action_Recognition_With_Shift_Graph_Convolutional_Network_CVPR_2020_paper.pdf

12 13 14 15 Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition
https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhang_Semantics-Guided_Neural_Networks_for_Efficient_Skeleton-Based_Human_Action_Recognition_CVPR_2020_paper.pdf

16 17 24 39 40 41 45 46 Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition
https://openaccess.thecvf.com/content/ICCV2023/papers/Lee_Hierarchically_Decomposed_Graph_Convolutional_Networks_for_Skeleton-Based_Action_Recognition_ICCV_2023_paper.pdf

18 19 20 21 22 23 25 Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition
https://openaccess.thecvf.com/content/ICCV2021/papers/Chen_Channel-Wise_Topology_Refinement_Graph_Convolution_for_Skeleton-Based_Action_Recognition_ICCV_2021_paper.pdf

26 27 28 29 30 [2106.15125] Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition
https://arxiv.org/abs/2106.15125

31 32 EfficientGCN: Faster Baselines for Action Recognition
https://www.emergentmind.com/papers/2106.15125

33 34 35 36 37 38 InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition
https://openaccess.thecvf.com/content/CVPR2022/papers/Chi_InfoGCN_Representation_Learning_for_Human_Skeleton-Based_Action_Recognition_CVPR_2022_paper.pdf

42 43 44 92 GitHub - Jho-Yonsei/HD-GCN: [ICCV 2023] Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition
https://github.com/Jho-Yonsei/HD-GCN

47 48 49 50 51 52 53 BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition - Microsoft Research
https://www.microsoft.com/en-us/research/publication/blockgcn-redefine-topology-awareness-for-skeleton-based-action-recognition/

54 55 56 57 58 59 60 Adaptive Hyper-Graph Convolution Network for Skeleton-based Human Action Recognition with Virtual Connections
https://openaccess.thecvf.com/content/ICCV2025/papers/Zhou_Adaptive_Hyper-Graph_Convolution_Network_for_Skeleton-based_Human_Action_Recognition_with_ICCV_2025_paper.pdf

61 62 86 87 88 89 GitHub - firework8/Awesome-Skeleton-based-Action-Recognition: A curated paper list of awesome skeleton-based action recognition.
https://github.com/firework8/Awesome-Skeleton-based-Action-Recognition

63 64 79 80 81 82 83 84 85 ecva.net
https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05796.pdf

65 Two-stream spatio-temporal GCN-transformer networks for skeleton ...
https://www.nature.com/articles/s41598-025-87752-8

66 (PDF) Two-stream spatio-temporal GCN-transformer networks for ...
https://www.researchgate.net/publication/388854107_Two-stream_spatio-temporal_GCN-transformer_networks_for_skeleton-based_action_recognition

67 Two-stream spatio-temporal GCN-transformer networks for skeleton ...
https://www.semanticscholar.org/paper/Two-stream-spatio-temporal-GCN-transformer-networks-Chen-Chen/d74841bf691a6bf47638668f6947133737127085

68 69 70 71 72 73 74 75 76 77 3Mformer: Multi-Order Multi-Mode Transformer for Skeletal Action Recognition
https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_3Mformer_Multi-Order_Multi-Mode_Transformer_for_Skeletal_Action_Recognition_CVPR_2023_paper.pdf

78 A GCN and Transformer complementary network for skeleton-based action recognition | Request PDF
https://www.researchgate.net/publication/385151146_A_GCN_and_Transformer_complementary_network_for_skeleton-based_action_recognition

90 Quo Vadis, Skeleton Action Recognition - NASA ADS
https://ui.adsabs.harvard.edu/abs/2020arXiv200702072G/abstract

91 Quo Vadis, Skeleton Action Recognition ? - CVIT, IIIT
http://cvit.iiit.ac.in/research/projects/cvit-projects/quo-vadis-skeleton-action-recognition