Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CD-Former: A Contextual Dynamic Transformer for Skeleton-Based Human Action Recognition

Romel Antonio Baldivia Calderón de la Barca
romel.baldivia@ucb.edu.bo
Universidad Católica Boliviana San Pablo
Daniel Ramiro Humberto Calderón Rocabado
Universidad Católica Boliviana San Pablo
Gabriel Antonio Rojas Silva
Universidad Católica Boliviana San Pablo
Miguel Angel Clavijo Quispe
Universidad Católica Boliviana San Pablo
Xavier Alexis Murillo Sánchez
Universidad Católica Boliviana San Pablo
Guillermo Enrique Manning Soria Galvarro
Universidad Católica Boliviana San Pablo
Carlos Helsner Menacho Guerra
Universidad Católica Boliviana San Pablo
Diego Cesar Calderón Iturricha
Universidad Católica Boliviana San Pablo

Article

Keywords: Skeleton-based action recognition, Transformer, Spatiotemporal modeling, Selfattention, Human action recognition

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11153560/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CD-Former: A Contextual Dynamic Transformer for Skeleton-Based Human Action Recognition

Romel Antonio Baldivia Calderón de la Barca[1,*], Daniel Ramiro Humberto Calderón Rocabado[1], Gabriel Antonio Rojas Silva[1], Miguel Angel Clavijo Quispe[1], Xavier Alexis Murillo Sánchez[1], Guillermo Enrique Manning Soria Galvarro[1], Carlos Helsner Menacho Guerra[1], Diego Cesar Calderón Iturricha[1]

[1]Centro de Investigación, Desarrollo e Innovación en Ingeniería Mecatrónica (CIDIMEC), Universidad Católica Boliviana "San Pablo", 14 de Septiembre Avenue No. 4807, corner of 2nd Street, Obrajes, La Paz, Bolivia

[*]Corresponding author: romel.baldivia@ucb.edu.bo

Abstract

Transformer-based methods have become increasingly relevant to skeleton-based action recognition because of their ability to model long-range spatiotemporal dependencies. However, their architectural and computational characteristics can vary substantially across different formulations. This work investigates these characteristics through CD-Former (Contextual Dynamic Transformer), a pure Transformer formulation evaluated under short temporal windows. The architecture applies standard multi-head self-attention to unified joint-frame tokens enriched with learnable temporal and joint identity embeddings, without auxiliary graph convolutional, convolutional, recurrent, hypergraph, or multibranch processing modules. Experiments on NTU RGB+D and NTU RGB+D 120 examine temporal windows of 16, 24, and 32 frames, with recognition results aggregated over four independent runs initialized with different random seeds. The evaluation combines benchmark recognition performance with analytical complexity estimates and model-only measurements of latency, throughput, and GPU memory usage on an NVIDIA Jetson TX2. These measurements provide a practical view of how recognition performance, computational demand, and inference behavior vary across the evaluated temporal configurations, including operating points relevant to resource-constrained platforms.

Keywords: Skeleton-based action recognition; Transformer; Spatiotemporal modeling; Self-attention; Human action recognition

1 Introduction

Human action recognition (HAR) has been extensively studied in recent years and plays a key role in many applications such as video surveillance, human-computer interaction (HCI), smart healthcare, and workplace safety monitoring [1, 2]. Among the representations used for HAR, skeleton-based action recognition has attracted growing attention because it represents human motion through body-joint geometry while reducing dependence on appearance cues such as background texture and illumination [3, 4]. In this paradigm, the human body is represented as a set of 2D or 3D joint coordinates evolving over time, providing a compact and semantically structured representation of motion. Skeleton sequences can be acquired directly using depth-sensing devices or estimated from RGB imagery through human pose-estimation pipelines [5, 6].

Early deep learning methods for skeleton HAR transformed skeleton sequences into pseudo-images and processed them using Convolutional Neural Networks (CNNs) to capture spatiotemporal patterns [7, 8]. Other approaches employed recurrent architectures to model the temporal evolution of skeleton sequences [9]. While these approaches demonstrated the effectiveness of deep learning

for skeleton-based action recognition, skeletal connectivity and inter-joint relationships were not always modeled explicitly, motivating subsequent methods that represent the human body through its underlying graph structure.

To model skeletal structure explicitly, Graph Convolutional Networks (GCNs) represent the human body as a spatiotemporal graph whose nodes correspond to joints and whose edges encode spatial or temporal relationships [10]. Early formulations such as ST-GCN employ predefined skeletal connectivity to aggregate information across neighboring joints and frames. Subsequent methods introduced adaptive or learnable graph relations to provide greater flexibility in modeling joint dependencies [11, 12]. Graph-based modeling has consequently become a widely used approach for skeleton action recognition, while also motivating the exploration of alternative formulations in which spatiotemporal relationships can be inferred directly from the input sequence without requiring a predefined skeletal graph.

Recent skeleton-based action recognition studies have explored self-attention mechanisms to model spatiotemporal dependencies among joints and frames [4, 13, 14]. The Neurocomputing review by Xin et al. [15] distinguishes pure Transformer, hybrid Transformer, and Transformer-style GCN formulations, and further organizes pure Transformer designs according to serial, parallel, or fused spatiotemporal processing. A broader architecture-oriented survey likewise identifies Transformers as one of the major deep-learning families used for 3D skeleton action recognition [16]. These approaches use representations based on joints, body parts, or spatiotemporal tokens and have been evaluated extensively on benchmarks such as NTU RGB+D and NTU RGB+D 120 [5]. At the same time, existing architectures differ substantially in their tokenization strategies, auxiliary modules, number of streams, and backbone designs, resulting in different computational characteristics and making direct architectural comparisons difficult. Within this design space, CD-Former is most naturally characterized as a pure spatiotemporal fusion Transformer formulation: sequence-dependent joint-frame relationships are modeled through standard self-attention without predefined graph priors or auxiliary graph convolutional, convolutional, recurrent, hypergraph, or multibranch modules.

Training configuration is also relevant when evaluating Transformer models under short temporal windows. In this work, AdamW optimization, cosine annealing, label smoothing, temporal token dropout, temporal jitter, classification-head reset, and progressive encoder unfreezing are used as part of the evaluated training protocol. These components are considered jointly as a training configuration rather than as individually validated methodological contributions, and no causal improvement is attributed to any single component.

This paper investigates a pure Transformer formulation for skeleton HAR under short temporal windows and characterizes its recognition performance and computational behavior. Skeleton sequences are represented as unified joint-frame tokens enriched with learnable temporal and anatomical identity information, while standard multi-head self-attention models sequence-dependent spatiotemporal relationships without requiring a predefined skeletal graph or auxiliary graph convolutional, convolutional, recurrent, hypergraph, or multibranch modules. The evaluation considers temporal windows of 16, 24, and 32 frames and reports recognition accuracy together with computational and runtime measurements, including GFLOPs, latency, throughput, and memory usage.

The main contributions of this work are summarized as follows:

- A pure Transformer formulation for skeleton HAR is evaluated under temporal windows of 16, 24, and 32 frames without auxiliary graph convolutional, convolutional, recurrent, hypergraph, or multibranch modules.
- Skeleton-specific contextual tokenization represents joint coordinates as contextual joint-frame tokens using learnable temporal and joint identity embeddings, while standard multi-head self-attention recomputes sequence-dependent spatiotemporal relationships for each input sequence.
- A documented training configuration for short temporal windows combines AdamW op-

timization, cosine annealing, label smoothing, temporal token dropout, temporal jitter, classification-head reset, and progressive encoder unfreezing. Because these components are evaluated jointly, the study does not attribute the observed differences to any single training component.

- The accuracy-computation trade-off is examined on NTU RGB+D and NTU RGB+D 120 across multiple temporal resolutions, reporting recognition performance together with GFLOPs and model-only runtime measurements, including latency, throughput, and memory usage on the evaluated hardware.

2 Related Work

Skeleton-based action recognition has evolved through several architectural paradigms, ranging from early CNN- and RNN-based models to graph-based, Transformer-based, hybrid, and hypergraph-based approaches. These methods differ in how they represent skeletal topology, temporal dependencies, and long-range joint interactions, as well as in the computational resources required for inference. This section reviews representative developments in these directions and examines the relationship between recognition performance, architectural complexity, and computational cost that motivates the present study.

2.1 Early Deep Learning Paradigms (CNN and RNN/LSTM)

Early research in skeleton HAR adapted conventional deep learning architectures to temporal sequences of 2D or 3D joint coordinates. Recurrent architectures captured motion dynamics by treating skeletons as temporally ordered sequences of joint vectors [9]. These methods established the foundations of end-to-end temporal modeling for skeleton action recognition. Their sequential computation, nevertheless, restricts parallel processing, while vectorized skeleton representations do not explicitly preserve the non-Euclidean topology of the human body.

A parallel research direction transformed skeleton coordinates into pseudo-images or structured feature maps, enabling the application of standard Convolutional Neural Networks (CNNs) for feature extraction [7]. Such representations proved effective for capturing local spatial and temporal patterns using mature convolutional architectures. The imposed grid structure, however, does not naturally correspond to skeletal connectivity and may limit the representation of articulated motion. These limitations encouraged the development of models capable of encoding the physical structure of the human body more explicitly.

2.2 Graph Convolutional Networks (GCNs)

The introduction of Graph Convolutional Networks (GCNs) marked a significant transition in skeleton action recognition. By representing the human body as a spatiotemporal graph, with joints modeled as nodes and bones represented as edges, GCN-based methods preserve the natural topology of the skeleton. The seminal ST-GCN [10] demonstrated that graph convolutions could jointly capture anatomical structure and temporal motion, establishing a strong foundation for subsequent research.

Later studies expanded this formulation through multiscale, adaptive, and hierarchical graph modeling. MS-G3D [17] introduced receptive fields operating at different spatial and temporal scales, whereas CTR-GCN [12] refined skeletal topology at the channel level to improve discriminative representation. Adaptive graph structures and multiple input streams were subsequently explored through two-stream adaptive GCNs [11] and multistream variants [18]. More recent architectures, including InfoGCN [19] and HD-GCN [20], further strengthened representation learning through information-theoretic objectives and hierarchical graph decomposition, respectively.

Recent graph-based methods use a wide range of mechanisms, including topology refinement, hierarchical decomposition, multigraph aggregation, and parallel input streams. These choices

provide different ways to encode skeletal structure and motion, but they also lead to heterogeneous computational and implementation profiles across methods. This diversity motivates complementary formulations that study spatiotemporal skeleton modeling without requiring an explicit graph-processing backbone.

2.3 Transformer Taxonomy, Pure Attention Models, and Hybrid Architectures

Self-attention offers an alternative mechanism for modeling relationships among joints and frames. Rather than restricting information exchange to predefined skeletal connections, attention weights are computed from input features and can represent long-range and sequence-dependent interactions. The Neurocomputing review by Xin et al. [15] distinguishes pure Transformer architectures from hybrid Transformer and Transformer-style GCN formulations. In that taxonomy, pure Transformer models are further divided into spatio-temporal serial, parallel, and fusion designs according to how spatial and temporal interactions are organized. Ren et al. [16] provide a broader architecture-oriented survey covering RNN-, CNN-, GCN-, and Transformer-based skeleton action recognition.

Pure attention-based skeleton modeling therefore predates the present work. DSTA-Net [21] constructs a decoupled spatial-temporal attention network without requiring a manually specified graph topology, while STST [22] develops specialized spatial and temporal Transformer processing. Other pure Transformer systems surveyed by Xin et al. include IIP-Former, STAR, TSTE, IGFormer, and START, illustrating that pure attention provides an established design family rather than a new category introduced by CD-Former [15].

Under that taxonomy, CD-Former is most naturally characterized as a *pure spatio-temporal fusion Transformer*: joint and temporal identities are embedded into unified joint-frame tokens and processed by the same standard multi-head self-attention layers, without an explicit separation into dedicated spatial and temporal Transformer branches. The contribution is therefore not the introduction of pure Transformer processing itself or a new attention operator. Instead, the present study examines contextual joint-frame tokenization within this pure-fusion formulation under short temporal windows and reports the associated recognition, analytical-computation, and model-only runtime behavior.

A broader line of work introduces more specialized Transformer mechanisms. ST-TR [4] models spatial and temporal dependencies through dedicated attention modules. FG-STFormer [23] combines focal and global processing, STAR-Transformer [13] uses spatiotemporal cross-attention, and STSA-Net [24] introduces segment-based attention. Frequency-aware approaches such as FreqMixFormer [25] extend temporal modeling with frequency-domain processing, while Skeleton MixFormer [26] and action-embedding approaches [27] introduce additional representation mechanisms. ActionFormer [28] combines Transformer and convolutional components, providing a hybrid alternative to attention-only processing.

Higher-order and partition-based formulations have likewise been explored. 3MFormer [29] integrates multi-order and multi-mode representations, AutoRegAd-HGFormer [30] combines Transformer processing with adaptive autoregressive hypergraph construction, and SkateFormer [31] partitions joints and frames according to skeletal-temporal relation types before applying attention. These methods demonstrate different ways of incorporating structural priors, specialized representations, or multiple processing paths. Because their inputs, stream counts, temporal resolutions, and profiling conventions differ, computational comparisons across these architectures require caution.

Recent Neurocomputing studies further illustrate the diversity of this design space. SPAR [32] reduces self-attention cost through switching partition strategies, STDM-Transformer [33] combines space-time partitioning with dual multi-scale Transformer processing, and LG-STFormer [34] integrates topology-constrained spatial Transformer processing with multiscale temporal convolution and joint-bone fusion. More recently, the multi-view transformer with hierarchical attention [35] combines multi-view geometric transformation, hierarchical attention, spatial and channel attention, an adaptive topology graph, and multiscale temporal convolution. HPI-GCN [36], although graph-

based rather than a pure Transformer, addresses practical inference through graph and temporal re-parameterization. Together, these works show that current skeleton-recognition research spans pure attention, hybrid, graph-based, and inference-oriented formulations rather than converging on a single architectural family.

More recent systematic reviews also show that the field has broadened beyond supervised backbone design toward self-supervised learning, few-shot and meta-learning, and multimodal representation learning [37]. CD-Former remains deliberately focused on the supervised single-modality setting; its scope is to characterize a comparatively simple pure Transformer operating regime rather than to address these broader learning paradigms.

2.4 Reduced-Compute Attention and Broader HAR Models

Computational cost is an important consideration in Transformer-based skeleton recognition because the self-attention interaction matrix grows quadratically with the number of input tokens. The Neurocomputing review by Xin et al. [15] identifies lightweight Transformer design for skeleton action recognition as a relevant future direction, specifically emphasizing the need to reduce computational requirements while retaining recognition performance. More broadly, recent architecture surveys likewise discuss computational cost as one of the practical dimensions separating skeleton-recognition designs [16].

Several studies address this problem directly through specialized efficient-attention or lightweight graph mechanisms. Qin et al. [38] investigate efficient self-attention for skeleton action recognition, while Zhang et al. [39] develop a lightweight graph convolutional design intended to reduce computational complexity. These approaches pursue architectural reductions that are different from the strategy evaluated in CD-Former.

The present work does *not* characterize CD-Former as a lightweight Transformer by parameter count: the model contains approximately 11.29M trainable parameters, with a small variation due to temporal-embedding length, and the 2048-dimensional feed-forward sublayers contribute substantially to this total. Instead, CD-Former provides reduced-computation operating points by shortening the temporal window and therefore reducing the number of joint-frame tokens processed by standard self-attention. The configurations with 16, 24, and 32 frames are analyzed explicitly in terms of recognition accuracy, analytical GFLOPs, latency, throughput, and measured memory usage.

Comparable efficiency-oriented developments have appeared in sensor-based HAR, where recent sequence models seek to reduce the cost of long-range temporal representation. ActiFormer [40] introduces sign-aware linear attention for wearable sensor signals, while TSA-Former [41] adopts a linear Transformer formulation based on Taylor-series attention. Parameter-efficient adaptation of time-series foundation models has also been explored through Sensor-Prompt Tuning [42], whereas DeepSenseMoE [43] introduces a mixture-of-experts adaptation module for few-shot HAR under heterogeneous sensor conditions.

Wearable time-series signals and skeleton-coordinate sequences correspond to different sensing modalities and learning conditions, so direct numerical comparisons between these method families are not appropriate. Their shared emphasis on controlling sequence-modeling cost nevertheless provides broader context for evaluating short-window operating points in skeleton-based HAR.

3 Methods

This work presents **CD-Former**, a pure Transformer architecture for skeleton-based Human Action Recognition (HAR). The model processes short spatiotemporal skeleton sequences through a unified attention-based pipeline and is evaluated under resource-constrained embedded GPU settings. Unlike hybrid approaches that integrate GCNs, CNNs, hypergraph reasoning, autoregressive modules, frequency-specific branches, or multistream motion encoders, the proposed architecture performs spatiotemporal modeling within a unified Transformer encoder.

The model operates on contextualized skeleton tokens. Each input joint coordinate is first projected into a latent feature space and enriched with learnable temporal and joint identity embeddings. Standard multi-head self-attention is then applied to the resulting spatiotemporal token sequence, allowing the relational weights to depend on the features of the current input while remaining conditioned by temporal order and anatomical joint identity. Consequently, the joint-frame interaction pattern is recomputed for each skeleton sequence rather than being imposed through a fixed skeletal adjacency structure.

In the name *Contextual Dynamic Transformer*, *contextual* refers to the explicit conditioning of each skeleton token on temporal position and anatomical joint identity, whereas *dynamic* refers to the input-dependent relational weights produced by standard self-attention. Because the queries and keys are computed from the current contextualized token sequence, the resulting joint-frame attention matrix changes with the input sequence. The term *dynamic* therefore does not denote a new attention operator, dynamic network topology, adaptive frame count, or an additional gating or routing mechanism. The attention computation itself remains standard scaled dot-product multi-head self-attention.

The proposed pipeline is composed of four main stages: (1) input preprocessing and contextual tokenization, (2) a stack of Post-LN Transformer encoder layers operating on contextualized skeleton tokens, (3) a global representation module that combines the CLS token with an average-pooled summary of all non-CLS tokens, and (4) a linear classification head that produces the final action prediction.

The complete CD-Former architecture and its main processing stages are illustrated in Fig. 1. Supplementary Fig. S1 further details the standard multi-head self-attention operation applied to the contextualized spatiotemporal skeleton tokens.

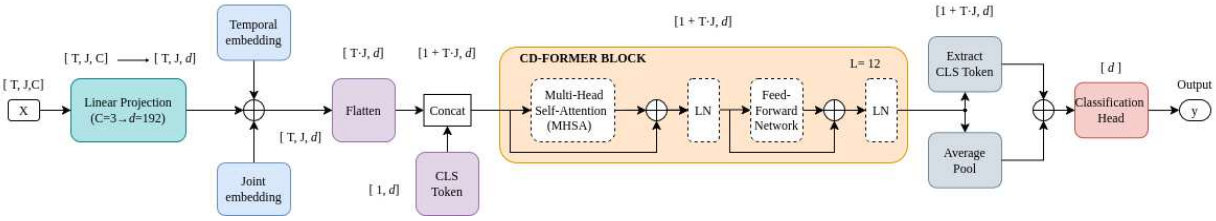


Figure 1: Overview of the proposed pipeline. Each input skeleton sequence is first projected into a latent feature space through a linear projection, producing joint-wise latent embeddings. Learnable temporal and joint identity embeddings are added to encode frame order and anatomical information. The resulting tensor is flattened into a unified spatiotemporal token sequence, and a learnable CLS token is prepended. The sequence is processed by a stack of $L = 12$ identical Post-LN Transformer encoder layers. In each layer, standard multi-head self-attention over contextualized skeleton tokens is followed by residual addition and layer normalization; the resulting representation is then processed by a feed-forward network, followed by a second residual addition and layer normalization. After the final encoder layer, the output representation combines the learned CLS representation with an average-pooled summary of all non-CLS tokens. The two representations are fused through element-wise addition and passed to a linear classification head to produce the final action prediction.

3.1 Input Preprocessing and Contextual Tokenization

Let $X \in \mathbb{R}^{T\times J\times C}$ denote the skeleton stream used by the model, with T frames, J joints, and C coordinate channels. The experiments use the preprocessed 3D skeleton annotation files for NTU RGB+D and NTU RGB+D 120 distributed through the PySKL project [44]. All action classes and all available samples in the processed protocol splits are retained, including recordings involving more than one actor. For multi-body recordings, the current input construction uses the

first indexed skeleton stream before temporal processing in order to preserve the fixed $T \times J \times C$ representation. This is an input-representation choice rather than a sample-selection rule: multi-person action samples remain part of both training and evaluation. Sequences longer than the target temporal window are center-cropped, whereas shorter sequences are padded by repeating the last available frame.

To reduce coordinate offset and scale variation, we apply **frame-wise z-score normalization**. For each frame t and coordinate channel c, the mean and standard deviation are computed across the J joints:

$$\tilde{x}_{t,j,c} = \frac{x_{t,j,c} - \mu_{t,c}}{\sigma_{t,c} + \epsilon}, \tag{1}$$

where $\mu_{t,c}$ and $\sigma_{t,c}$ denote the frame-wise joint mean and standard deviation, respectively. This normalization emphasizes relative articulated pose and reduces sensitivity to absolute coordinate offsets. A consequence is that part of the global body-translation signal is suppressed; therefore, actions whose discrimination depends strongly on absolute inter-frame displacement may be more challenging under this preprocessing choice.

Each normalized joint vector $\tilde{x}_{t,j}$ is independently projected into a d-dimensional latent space:

$$z_{t,j} = \tilde{x}_{t,j} W_{\text{proj}} + b_{\text{proj}}, \tag{2}$$

yielding $N = T \times J$ joint-level embeddings. Because self-attention does not inherently encode temporal order or anatomical joint identity, these identities must be explicitly represented. We introduce three learnable components:

- **Temporal embedding** E_t, encoding the frame index.
- **Joint embedding** E_j, encoding the anatomical identity of each joint.
- **CLS token** h_{cls}, which aggregates global spatiotemporal context.

For each joint-frame location, the contextualized representation is defined as

$$s_{t,j} = z_{t,j} + e_t^{\text{temp}} + e_j^{\text{joint}}, \tag{3}$$

where $e_t^{\text{temp}} \in \mathbb{R}^d$ denotes the temporal embedding and $e_j^{\text{joint}} \in \mathbb{R}^d$ denotes the joint identity embedding.

The contextualized joint-frame representations are flattened into a unified sequence, and the learnable classification token is prepended:

$$S^{(0)} = [\, h_{\text{cls}};\, s_{1,1}, \ldots, s_{T,J} \,] \in \mathbb{R}^{(N+1)\times d}. \tag{4}$$

This form of *contextual tokenization* is central to the proposed design. Temporal progression and anatomical identity are embedded directly into the token representation, allowing the self-attention layers to model human motion without requiring predefined skeletal adjacency matrices or graph convolutional priors.

3.2 Computational Complexity Analysis

Let $N = TJ$ denote the number of joint-frame tokens generated from a skeleton sequence with T frames and J joints. After prepending the learnable CLS token, the complete sequence length is

$$M = N + 1 = TJ + 1. \tag{5}$$

In multi-head self-attention, the attention matrix therefore satisfies

$$QK^\top \in \mathbb{R}^{M\times M}. \tag{6}$$

The quadratic attention term has time and memory complexity proportional to

$$\mathcal{O}(M^2 d) = \mathcal{O}((TJ+1)^2 d), \tag{7}$$

where d is the embedding dimension.

For the computational-cost analysis, the principal multiply-accumulate operations of one Transformer encoder layer are counted explicitly. The query, key, value, and output projections require

$$4Md^2 \tag{8}$$

MACs. The two matrix multiplications associated with self-attention, $QK^\top$ and the multiplication of the attention matrix by V, require

$$2M^2 d \tag{9}$$

MACs, while the two linear transformations of the feed-forward network require

$$2Md\,d_{\text{ff}} \tag{10}$$

MACs. Consequently, the principal matrix-operation cost of one encoder layer is

$$\text{MACs}_{\text{layer}} = 4Md^2 + 2M^2 d + 2Md\,d_{\text{ff}}. \tag{11}$$

Including the input projection and final classification layer, the total analytical MAC count is

$$\text{MACs}_{\text{total}} = TJCd + L\left(4Md^2 + 2M^2 d + 2Md\,d_{\text{ff}}\right) + dK, \tag{12}$$

where C is the input coordinate dimension, L is the number of Transformer encoder layers, and K is the number of output classes.

The computational values reported in this work use the convention that one multiply-accumulate operation corresponds to two floating-point operations:

$$\text{FLOPs}_{\text{total}} = 2\,\text{MACs}_{\text{total}}. \tag{13}$$

For the evaluated architecture, $d = 192$, $d_{\text{ff}} = 2048$, $L = 12$, and $J = 25$. Under this counting convention, the configurations with 16, 24, and 32 frames require 10.47, 16.80, and 23.87 GFLOPs, respectively. The calculation includes the input projection, Q/K/V projections, attention output projection, both attention matrix multiplications, feed-forward network, and final classification layer. Element-wise operations such as activation functions, normalization, Softmax, and bias additions are not included in this principal matrix-operation count.

The quadratic component is the term most directly affected by temporal resolution. Reducing T decreases the number of joint-frame tokens and therefore reduces both the linear token-dependent terms and the quadratic self-attention term. This trade-off motivates the evaluation of CD-Former with $T \in \{16, 24, 32\}$ frames.

3.3 Transformer Encoder over Contextualized Skeleton Tokens

The encoder consists of $L = 12$ identical Post-LN layers. Each layer comprises: (1) standard multi-head self-attention applied to contextualized skeleton tokens, (2) a feed-forward network (FFN), and (3) two residual connections followed by layer normalization.

3.3.1 Standard Multi-Head Self-Attention

Given input tokens S, the standard Q/K/V projections are computed as

$$Q = SW_Q, \qquad K = SW_K, \qquad V = SW_V. \tag{14}$$

Scaled dot-product attention is then computed independently for each head:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{15}$$

where d_k denotes the dimensionality of the query and key vectors in each attention head.

For a contextualized input sequence S, the corresponding attention-weight matrix can be written as

$$A(S) = \text{softmax}\left(\frac{(SW_Q)(SW_K)^\top}{\sqrt{d_k}}\right). \tag{16}$$

Since $A(S)$ is computed from the current sequence, its relational pattern is input-dependent and can vary across action instances even though the attention operator and network parameters remain fixed. This is the precise sense in which CD-Former uses the term *dynamic*. Unlike conventional applications of self-attention to generic token sequences, the proposed model injects temporal and joint identity embeddings into the skeleton tokens before the attention operation. Consequently, the similarity scores represented by $QK^\top$ depend jointly on the current token content, frame position, and anatomical joint identity.

The resulting attention relations are recomputed for each input sequence. In the evaluated architecture, this unified attention pathway is used without predefined skeletal adjacency matrices, graph convolutional operations, or auxiliary structural modules.

3.3.2 Feed-Forward Network

Each attention output is processed by a position-wise feed-forward network:

$$\text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2, \tag{17}$$

where σ is the ReLU activation function. The FFN provides nonlinear feature transformation within each Transformer encoder layer. Inference cost is reduced primarily through the reduced temporal input length and the resulting decrease in the number of spatiotemporal tokens.

3.3.3 Layer Composition

Each encoder layer follows a Post-LN Transformer structure, where residual additions are performed before layer normalization. For encoder layer $\ell \in \{1, \ldots, L\}$, with $L = 12$, the composition is

$$U^{(\ell)} = \text{LayerNorm}\left(S^{(\ell-1)} + \text{MHSA}\left(S^{(\ell-1)}\right)\right), \tag{18}$$

$$S^{(\ell)} = \text{LayerNorm}\left(U^{(\ell)} + \text{FFN}\left(U^{(\ell)}\right)\right). \tag{19}$$

The output $S^{(\ell)}$ is subsequently used as the input to the following encoder layer.

3.4 Global Representation and Classification

After L encoder layers, let the final output sequence be

$$S^{(L)} = [\, h_{\text{cls}}^{(L)}; h_1^{(L)}, \ldots, h_N^{(L)} \,]. \tag{20}$$

Average pooling is applied exclusively to the N non-CLS tokens:

$$h_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} h_i^{(L)}. \tag{21}$$

The global representation is obtained through element-wise addition:

$$h = h_{\text{cls}}^{(L)} + h_{\text{avg}} \in \mathbb{R}^d. \tag{22}$$

The CLS token serves as a learned global representation, while average pooling summarizes the non-CLS joint-frame tokens. Element-wise fusion maintains the original embedding dimension d before the linear classification head.

The linear head produces class logits:

$$z = hW_{\text{head}} + b_{\text{head}}, \qquad W_{\text{head}} \in \mathbb{R}^{d \times K}, \tag{23}$$

where K denotes the number of action classes. During inference, the Softmax function converts the logits into class probabilities.

3.5 Training Configuration

The evaluated training configuration combines: (A) AdamW optimization, (B) cosine annealing scheduling, (C) label smoothing regularization, (D) temporal perturbations based on temporal token dropout and temporal jitter, (E) classification-head reset, and (F) encoder freezing followed by end-to-end unfreezing.

The hyperparameters reported in Supplementary Table S1 correspond to the 32-frame configuration whose mean X-Set accuracy over four independent runs is 92.96%, and serve as the reference setting described in this section. The optimization and regularization components are evaluated jointly; therefore, the reported configuration-level comparisons do not isolate the causal contribution of any individual component. Controlled component-wise ablations remain an appropriate direction for future work.

Checkpoint selection and reporting scope For the reported runs, early stopping and best-checkpoint selection monitored Top-1 accuracy on the protocol-specific partition identified as `xsub_val` or `xset_val` in the PySKL annotation file [44]. In the NTU RGB+D 120 protocol, these partitions correspond to the official evaluation partitions rather than to an independently held-out validation subset. Consequently, the reported accuracies are checkpoint-selected evaluation values and should not be interpreted as strictly untouched-test estimates. This limitation is stated explicitly to distinguish the reported experimental procedure from protocols that reserve an independent validation subset for model selection.

3.5.1 AdamW with Decoupled Weight Decay

AdamW decouples weight decay from the adaptive gradient update, providing direct control over parameter regularization. The parameter update rule is

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta \lambda \theta_t. \tag{24}$$

AdamW is used in the reported training configuration to apply decoupled weight decay during optimization.

3.5.2 Cosine Annealing Learning Rate Scheduling

The learning rate follows a cosine annealing schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \frac{\pi t}{T_{\max}}\right), \tag{25}$$

where $\eta_{\max}$ is the initial learning rate, $\eta_{\min}$ is the minimum learning rate, t is the current epoch, and $T_{\max}$ is the maximum number of training epochs. The cosine schedule provides smooth learning-rate decay and is employed to avoid abrupt optimization changes.

3.5.3 Label Smoothing

To mitigate overconfidence among visually similar actions, we employ label smoothing. Let $y_k \in \{0, 1\}$ denote the k-th component of the one-hot target vector. The smoothed target distribution is defined as

$$y_k^{\mathrm{LS}} = (1 - \varepsilon) y_k + \frac{\varepsilon}{K}, \qquad k = 1, \ldots, K, \tag{26}$$

where K denotes the number of action classes and ε controls the smoothing strength.

For the ground-truth class, $y_k = 1$, yielding $1 - \varepsilon + \varepsilon/K$. For all remaining classes, $y_k = 0$, yielding ε/K. This formulation reduces overconfident predictions, particularly among fine-grained motion categories.

3.5.4 Temporal Perturbation Augmentations: Temporal Token Dropout and Temporal Jitter

Skeleton sequences may contain missing or unreliable temporal observations and variations in temporal alignment. The reported training configuration therefore includes two temporal perturbation strategies.

Temporal Token Dropout Let $X_t \in \mathbb{R}^{J \times C}$ denote the complete skeleton observation at frame t. During training, a subset $\mathcal{D} \subseteq \{1, \ldots, T\}$ of temporal positions is selected, and all joint observations associated with those positions are jointly masked. The perturbed sequence is defined as

$$X_t^{\mathrm{drop}} = m_t X_t, \qquad t = 1, \ldots, T, \tag{27}$$

where $m_t = 0$ for $t \in \mathcal{D}$ and $m_t = 1$ otherwise. The number of masked temporal positions is

$$|\mathcal{D}| = \lfloor p_{\mathrm{drop}} T \rfloor. \tag{28}$$

This perturbation removes complete temporal observations during training while leaving the architecture and inference-time input format unchanged.

Temporal Jitter To simulate variation in the temporal alignment of the sampled window, a random circular shift is applied to the skeleton sequence:

$$X_t^{\mathrm{jitter}} = X_{1+((t-1+\delta) \bmod T)}, \qquad \delta \sim \mathrm{Unif}\{-\tau, \ldots, \tau\}. \tag{29}$$

This perturbation exposes the model during training to small changes in the temporal alignment of the input sequence.

3.5.5 Classification Head Reset

The evaluated training procedure reinitializes the classification-head weights W_{head} using Xavier uniform initialization after an initial training stage, while preserving the encoder weights. This is implemented as a two-stage procedure: the complete model is first trained and a checkpoint is saved; the classification head is then reinitialized, and training resumes from that checkpoint with the encoder weights preserved.

Formally, let $\mathcal{E}$ denote the encoder and $\mathcal{H}$ the classification head. At the reset step,

$$W_{\text{head}} \leftarrow \text{XavierUniform}\left(W_{\text{head}}\right), \qquad b_{\text{head}} \leftarrow \mathbf{0}_K. \tag{30}$$

This two-stage strategy does not modify the architecture, parameter count, or GFLOPs. Under the 16-frame setting, the complete optimized training configuration yields gains of up to 9.84 percentage points relative to the corresponding baseline configuration. Since head reinitialization, regularization, and encoder unfreezing are modified jointly, this configuration-level comparison does not isolate the individual contribution of the classification-head reset.

3.5.6 Progressive Encoder Unfreezing

To reduce abrupt changes in the previously learned encoder representations during the head-reset procedure, the first N_{freeze} encoder layers are frozen immediately after head reinitialization. After ΔE training epochs, all encoder weights are unfrozen and training continues end-to-end. Formally,

$$\mathcal{E}_{1:N_{\text{freeze}}} \leftarrow \text{frozen for } \Delta E \text{ epochs after reset}, \tag{31}$$

where $\mathcal{E}_{1:N_{\text{freeze}}}$ denotes the first N_{freeze} encoder layers. This procedure allows the newly initialized head to adapt to fixed encoder representations before joint fine-tuning begins. For the best-performing configuration identified in our experiments, $N_{\text{freeze}} = 12$ and $\Delta E = 15$ were used, as reported in Supplementary Table S1.

Together, these strategies constitute the evaluated training configuration while preserving the architecture, parameter count, and inference cost; their individual effects are not isolated in this study.

3.6 Ethical considerations

This study constitutes a secondary analysis of previously collected third-party benchmark data and did not involve the recruitment of participants, collection of new human-subject data, or execution of new experimental protocols by the authors. The experiments use preprocessed 3D skeleton annotations for NTU RGB+D and NTU RGB+D 120 distributed through the PySKL project and derived from the original datasets provided by the ROSE Lab at Nanyang Technological University. According to the official ROSE Lab dataset documentation, informed consent was obtained from all subjects whose data were released through the dataset repository. All methods in the present study were carried out in accordance with relevant guidelines and regulations applicable to the use of these datasets. No additional human-participant experiments were conducted as part of the present study.

3.7 Use of artificial intelligence in manuscript preparation

ChatGPT (OpenAI) was used solely to assist with language refinement and editorial organization of the manuscript. It was not used to generate experimental data, perform experiments, conduct statistical analyses, or determine the scientific results or conclusions. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the content of the manuscript.

4 Experiments

4.1 Datasets

4.1.1 NTU RGB+D

NTU RGB+D is one of the most widely used datasets for evaluating skeleton human action recognition methods [45]. It contains 56,880 RGB+D video samples covering 60 action categories.

All sequences were recorded using the Microsoft Kinect v2 depth sensor from three horizontal viewpoints ($-45°$, $0°$, $45°$), and actions were performed by 40 subjects aged between 10 and 35 years. Each frame provides 3D spatial coordinates for 25 major human joints, and each clip contains at most two subjects.

These protocols evaluate performance under subject variation and viewpoint changes, respectively. NTU RGB+D provides an additional 60-class benchmark for evaluating the same CD-Former formulation under the established Cross-Subject and Cross-View protocols.

Table 1: Comparison with representative methods on NTU RGB+D (60 classes). CD-Former results are mean ± standard deviation over four independent runs initialized with different random seeds.

Model	**X-Sub**	**X-View**	**Params (M)**	**Frames**
CTR-GCN (4-ensemble) [12]	92.4	96.8	5.84	64
3M-Former (2-branch) [29]	94.8	98.7	4.37	64
FG-STFormer [23]	92.6	96.7	–	128
CD-Former (16f)	80.12 ± 0.14	85.33 ± 0.12	11.28	16
CD-Former (24f)	87.06 ± 0.11	92.00 ± 0.12	11.29	24
CD-Former (32f)	89.91 ± 0.10	94.45 ± 0.12	11.29	32

Following the standard evaluation protocols, we report results under the two official benchmarks defined in the original dataset paper [45]:

- **Cross-Subject (X-Sub):** 40 subjects are split into mutually exclusive training and testing groups. The training set contains 40,320 sequences and the testing set contains 16,560 sequences.
- **Cross-View (X-View):** training samples are captured from cameras at $0°$ and $45°$, and testing samples are captured from the $-45°$ viewpoint.

4.1.2 NTU RGB+D 120

NTU RGB+D 120 extends NTU RGB+D to 114,480 samples, 120 action classes, and 106 subjects [5]. The recordings span 32 different setups, where each setup corresponds to variations in scene layout, background, and camera placement. Joint annotations remain consistent with NTU RGB+D, providing 25 3D joints per subject per frame.

Two official benchmarks are recommended:

- **Cross-Subject (X-Sub):** the 106 subjects are divided into training and testing groups following the protocol in the original paper.
- **Cross-Setup (X-Set):** samples recorded in even numbered setups are used for training, and those in odd numbered setups are used for testing.

Compared with NTU RGB+D, NTU RGB+D 120 expands the benchmark in the number of action classes, subjects, acquisition setups, and recorded conditions [5]. The two official protocols therefore provide a broader setting for evaluating recognition under subject and setup variation.

The same CD-Former formulation is evaluated on NTU RGB+D 120 without introducing predefined graph priors, temporal-convolution modules, or multistream motion encoders.

To document the behavior of the combined training configuration, which integrates AdamW, label smoothing, cosine annealing, temporal token dropout, temporal jitter, classification-head reset, and progressive encoder unfreezing, we conduct a configuration-level analysis across three sequence lengths: $T \in \{16, 24, 32\}$ frames. Each pair of baseline and optimized models shares the same architecture, parameter count, and GFLOPs. Therefore, the observed differences cannot be

attributed to increased model capacity. However, because several optimization and regularization components are modified jointly, this analysis does not isolate their individual contributions.

We report results under both Cross-Subject (X-Sub) and Cross-Setup (X-Set) protocols of NTU RGB+D 120.

The reported results cover all action classes and all available samples in the processed X-Sub and X-Set protocol splits, including two-person interaction recordings. For multi-body samples, CD-Former constructs its input from the first indexed skeleton stream and does not fuse an additional body stream, thereby maintaining the same $T \times J \times C$ input structure across the dataset. Consequently, this preprocessing choice does not remove two-person actions from the benchmark; it only defines how multi-body recordings are represented at the model input.

4.1.3 Training-configuration analysis across temporal windows

The configuration-level comparison is summarized across the three evaluated temporal windows in Table 2. Each baseline and optimized pair shares the same architecture, parameter count, and analytical GFLOPs. Multiple optimization and regularization components differ jointly; therefore, the comparison does not isolate the contribution of any single component.

Table 2: Training-configuration comparison across 16, 24, and 32 frames. Optimized CD-Former results are reported as mean ± standard deviation over four independent runs initialized with different random seeds; baseline values are point estimates from the original baseline experiments.

Frames	Configuration	X-Sub (%)	X-Set (%)	Gain (pp)
16	Baseline	75.84	76.04	–
16	CD-Former	$\mathbf{84.48 \pm 0.12}$	$\mathbf{85.88 \pm 0.12}$	+8.64 / +9.84
24	Baseline	88.04	87.52	–
24	CD-Former	$\mathbf{88.22 \pm 0.10}$	$\mathbf{92.00 \pm 0.13}$	+0.18 / +4.48
32	Baseline	**90.52**	90.10	–
32	CD-Former	89.28 ± 0.10	$\mathbf{92.96 \pm 0.12}$	-1.24 / +2.86

At 16 frames, the combined training configuration produces higher X-Sub and X-Set accuracy than the corresponding baseline. At 24 frames, the X-Sub change is small while the X-Set improvement is larger. At 32 frames, the optimized configuration achieves the highest X-Set accuracy among the evaluated temporal resolutions, while its mean X-Sub accuracy is 1.24 percentage points below the baseline point estimate. These protocol-dependent differences are interpreted at the configuration level rather than as evidence for the effect of any individual optimization or regularization component.

Each CD-Former configuration was trained in four independent runs initialized with different random seeds, while keeping the data partitions, architecture, optimization settings, and evaluation protocol unchanged. The corresponding mean ± standard deviation is reported in the configuration tables above and in Tables 1 and 3. The standard deviation therefore reflects run-to-run training variability for a fixed configuration rather than variation across the operating points at 16, 24, and 32 frames. The observed standard deviations are small (0.10–0.14 percentage points across the reported benchmark settings); they are presented descriptively and are not attributed to any single optimization or regularization component.

4.1.4 Overall Discussion

Across all frame lengths, CD-Former exhibits a clear trade-off between temporal resolution, recognition accuracy, and computational cost. Three key patterns emerge:

- **Temporal Context:** Higher recognition accuracy is observed with longer temporal windows in the evaluated configurations. The 32-frame configuration provides the highest X-Set accuracy among the three temporal resolutions considered.

- **Reduced Computation at Short Windows:** The 16-frame configuration uses the fewest spatiotemporal tokens and provides the lowest analytical cost and measured Jetson TX2 latency among the evaluated CD-Former settings.
- **Protocol-Dependent Behavior:** The optimized configuration produces larger gains under X-Set than under X-Sub, showing that the training configuration affects the two official evaluation protocols differently.

These results support the use of CD-Former as a pure Transformer framework for skeleton action recognition. Rather than relying on hybrid graph, convolutional, recurrent, or multibranch modules, CD-Former uses contextual spatiotemporal tokenization and standard self-attention to provide several operating points across recognition performance and computational cost.

4.2 Comparison with Representative Methods

Table 3 presents a comparison between the proposed model and representative methods on NTU RGB+D 120 under the X-Sub and X-Set evaluation protocols.

Table 3: Comparison with representative methods on NTU RGB+D 120. CD-Former results are mean ± standard deviation over four independent runs initialized with different random seeds. Ensemble or branch configurations are identified in the model label when the cited result corresponds to such a configuration. Accuracy, parameter count, frame count, and GFLOP entries for competing methods are reported from the cited sources when available; unavailable or unverified complexity entries are marked "–". CD-Former GFLOPs are analytically calculated under the two-FLOPs-per-MAC convention described in Section 3.2.

Model	**X-Sub (%)**	**X-Set (%)**	**Params (M)**	**GFLOPs**	**Frames**
3M-Former (2-branch) [29]	92.00	93.80	4.37	58.45	64
AutoRegAd-HGFormer [30]	91.02	92.42	3.20	15.40	64
HD-GCN (6-ensemble) [20]	90.10	91.60	–	–	64
InfoGCN (6-ensemble) [19]	89.80	91.20	–	–	64
FG-STFormer [23]	89.00	90.60	–	–	128
CTR-GCN (4-ensemble) [12]	88.90	90.60	5.84	7.88	64
DGAN [46]	89.70	90.90	–	–	64
STSA-Net [24]	88.50	90.70	–	–	64
MVTHA [35]	90.80	91.80	1.97	3.02	64
CD-Former (16f)	84.48 ± 0.12	85.88 ± 0.12	11.28	10.47	16
CD-Former (24f)	88.22 ± 0.10	92.00 ± 0.13	11.29	16.80	24
CD-Former (32f)	89.28 ± 0.10	92.96 ± 0.12	11.29	23.87	32

The compared methods were originally evaluated under heterogeneous experimental conditions, including different temporal input lengths, input representations, preprocessing pipelines, training strategies, implementations, numerical precision, and hardware platforms. In addition, the reported CD-Former checkpoints were selected using the protocol-specific PySKL evaluation partitions, as documented above, whereas many benchmark studies reserve the official test set for final evaluation only. Consequently, the accuracy, parameter count, and GFLOP values should be interpreted as contextual reference indicators rather than as results obtained under a fully controlled common environment.

Cross-paper computational values require particular caution. Different studies may employ different profiling tools, frame counts, input modalities, stream counts, and conventions for counting multiply-accumulate operations. The CD-Former values reported here follow the explicit analytical convention defined in Section 3.2 and include both self-attention matrix multiplications. They should therefore not be interpreted as strictly normalized against every computational value reported by competing methods.

The proposed model achieves 89.28 ± 0.10% on X-Sub and 92.96 ± 0.12% on X-Set using 32 frames. The configurations with 16, 24, and 32 frames provide distinct operating points for analyzing the relationship between temporal resolution, recognition performance, and computational cost. Several compared methods report higher accuracy under specific protocols, while generally operating with different temporal, architectural, and implementation settings.

The comparison with the 2026 multi-view transformer with hierarchical attention [35] further illustrates a difference in design objective. That model reports 90.8% X-Sub and 91.8% X-Set with 1.97M parameters and 3.02 GFLOPs using 64 input frames, while combining multi-view processing, hierarchical attention, adaptive topology modeling, and multiscale temporal convolution. CD-Former uses a shorter 32-frame input and reports 89.28 ± 0.10% X-Sub and 92.96 ± 0.12% X-Set, but with a larger parameter count and analytical operation cost. Accordingly, CD-Former is not positioned as the minimum-cost architecture; the comparison instead highlights its use of a comparatively direct pure Transformer formulation without auxiliary graph or temporal-convolution modules.

For additional historical context within the pure Transformer family, the Neurocomputing survey by Xin et al. [15] reports NTU RGB+D 120 results of 86.6%/89.0% for DSTA-Net and 88.4%/89.7% for IIP-Former under X-Sub/X-Set. The same survey lists START, a cross-modal method placed in its pure-fusion category, at 90.3%/92.7%. The 32-frame CD-Former mean result of 89.28%/92.96% therefore falls within the upper performance range of the pure Transformer entries summarized in that historical survey. This observation is used only as contextual evidence rather than as a formal ranking because the surveyed methods differ in stream configuration, input modality, representation, and other experimental details, and the survey predates more recent architectures.

CD-Former uses approximately 11.29M parameters (11.28M at 16 frames and 11.29M after rounding at 24 and 32 frames), which is higher than several of the compared methods; therefore, the present work does not claim parameter compactness. The 2048-dimensional feed-forward sublayer reported in Supplementary Table S1 accounts for a substantial fraction of this parameter count. Accordingly, the deployment profile is characterized jointly through parameter count, temporal input length, analytical GFLOPs, latency, and measured VRAM rather than through parameter count alone.

4.3 Accuracy-Computation Landscape

In addition to reporting recognition accuracy and computational cost independently, we examine their relationship across the three temporal configurations of CD-Former. Fig. 2 places the configurations with 16, 24, and 32 frames within a broader accuracy-computation landscape on NTU RGB+D 120 under the X-Set protocol.

The representative operating points visualized in the figure include CTR-GCN [12], Block-GCN [47], AutoRegAd-HGFormer [30], PoseConv3D [48], RA-GCNv2 [49], 2s-AGCN [11], MS-G3D [17], DSTA-Net [21], and 3M-Former [29]. Their computational values are used only to provide a broad literature context because the corresponding papers and reproductions may follow different operation-counting conventions.

As shown in Fig. 2, increasing the temporal input from 16 to 24 frames raises the computational cost from 10.47 to 16.80 GFLOPs while improving X-Set accuracy from a mean of 85.88% to 92.00%. Increasing the input further to 32 frames results in 23.87 GFLOPs and a mean X-Set accuracy of 92.96%. Thus, the transition from 16 to 24 frames produces a 6.12 percentage-point increase in X-Set accuracy for an additional 6.33 GFLOPs, whereas the transition from 24 to 32 frames produces a further 0.96 percentage-point increase for an additional 7.07 GFLOPs.

The 24-frame configuration therefore represents an intermediate operating point between the lower-cost 16-frame configuration and the highest-accuracy 32-frame configuration. Within the evaluated operating points, the increase from 24 to 32 frames yields a smaller accuracy gain than the increase from 16 to 24 frames, despite a larger additional analytical cost.

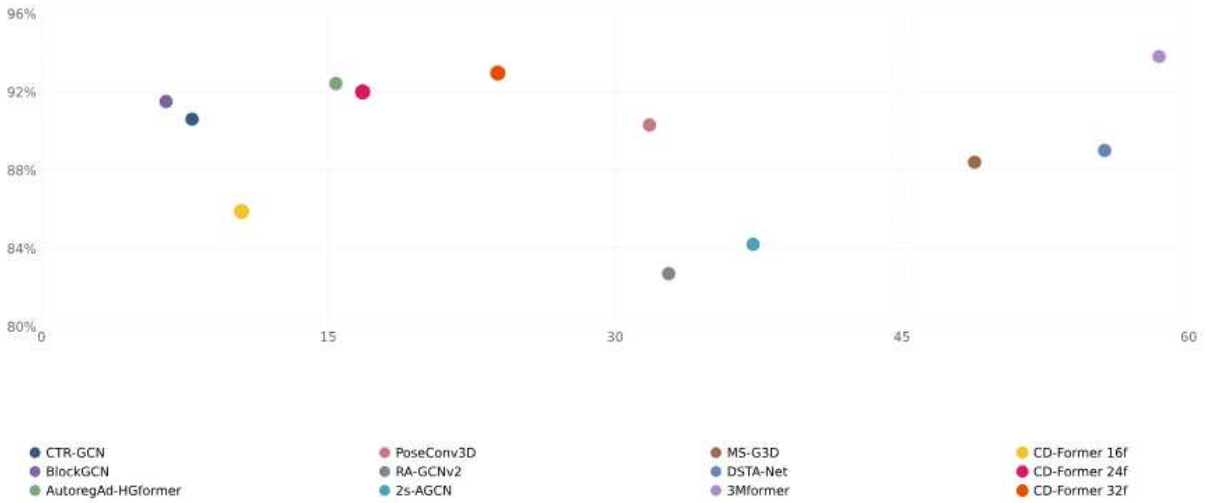


Figure 2: X-Set Top-1 accuracy versus reported computational cost on NTU RGB+D 120 for representative skeleton-based action recognition methods. CD-Former is shown at operating points of 16, 24, and 32 frames, requiring 10.47, 16.80, and 23.87 GFLOPs, respectively. CD-Former accuracy coordinates use the means over four independent runs reported in Table 3; standard deviations are omitted from the plot for readability. CD-Former computational costs are analytically calculated using the principal Transformer matrix operations under a two-FLOPs-per-MAC convention. Computational values for competing methods are taken from their respective publications or published reproductions and may follow different profiling procedures, input configurations, stream counts, and counting conventions. Consequently, the figure illustrates the broader accuracy-computation landscape rather than a strictly normalized efficiency ranking.

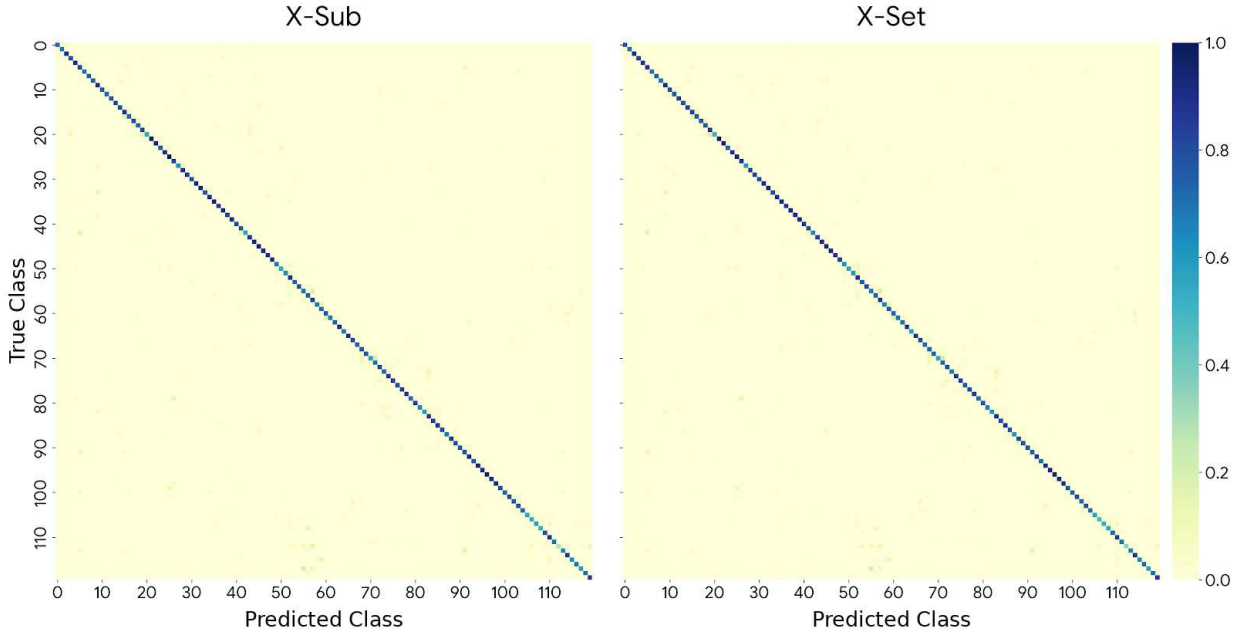
The broader comparison also shows that CD-Former does not occupy the lowest-computation region of the accuracy-cost space. Several lightweight graph-based architectures operate at lower reported GFLOP values. Conversely, representative methods such as DSTA-Net and 3M-Former occupy substantially higher reported computational regions, illustrating that the 32-frame CD-Former configuration lies within a broader range of computational profiles used for skeleton action recognition.

Importantly, the position of a method in Fig. 2 should not be interpreted as a fully controlled efficiency ranking. Cross-paper GFLOP values may differ because of profiling software, input representation, frame length, number of streams, precision, implementation details, and the convention used to count multiply-accumulate operations. The purpose of the figure is therefore to contextualize the three CD-Former operating points rather than to establish a universal computational ordering across architectures.

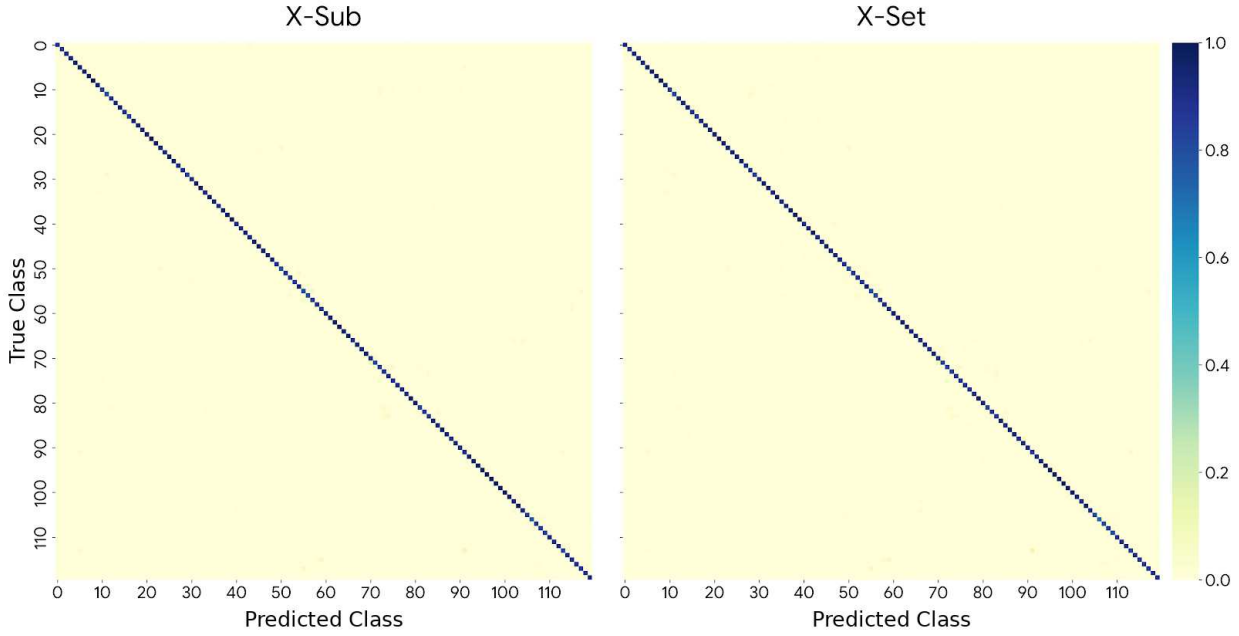
4.4 Class-Wise Recognition Analysis

To complement the aggregate Top-1 accuracy and computational-cost analysis, we examine the class-wise recognition behavior of CD-Former at the two endpoints of the evaluated temporal range. Specifically, Fig. 3 presents normalized confusion matrices for the 16-frame and 32-frame configurations under both the X-Sub and X-Set protocols of NTU RGB+D 120. The intermediate 24-frame configuration is retained in the quantitative tables and in the accuracy-computation analysis, while the two endpoint configurations are shown here to provide a clearer visualization of how class-wise recognition changes as the available temporal context increases.

The 16-frame configuration represents the lowest-computation operating point evaluated in this study. It requires 10.47 GFLOPs and obtains 84.48 ± 0.12% X-Sub and 85.88 ± 0.12% X-Set accuracy on NTU RGB+D 120. As shown in Fig. 3(a), the matrices retain a clear overall diagonal



(a) 16-frame configuration



(b) 32-frame configuration

Figure 3: Normalized confusion matrices of CD-Former on NTU RGB+D 120 for the two endpoint temporal configurations considered in this work: (a) 16 frames and (b) 32 frames. For each configuration, the left panel corresponds to X-Sub and the right panel to X-Set. Rows represent true action classes and columns represent predicted action classes. Each row is normalized independently, such that diagonal values correspond to class-wise recall. Several classes exhibiting pronounced confusion under the 16-frame input show substantially stronger diagonal responses at 32 frames, including action pairs whose distinction depends on the temporal progression of similar skeletal configurations.

structure, but several action categories exhibit pronounced off-diagonal responses. These errors indicate that a short temporal observation can be insufficient for distinguishing actions that contain similar intermediate poses or closely related movement patterns.

A representative case is *take off a hat/cap* (A21), which exhibits confusion with *put on a hat/cap* (A20) in the 16-frame configuration. Although these actions have opposite semantic directions, portions of their skeletal trajectories can contain highly similar joint configurations when only a short temporal segment is observed. The corresponding diagonal response becomes substantially stronger at 32 frames and the confusion between the two actions is markedly reduced.

A similar behavior is visible for *take off a bag* (A88) and *put on a bag* (A87). Under the shorter temporal window, the two categories exhibit greater confusion, whereas the 32-frame configuration provides a considerably clearer separation. These inverse-action pairs are particularly informative because their distinction is not defined solely by an isolated body pose, but also by the temporal ordering and progression of the movement.

Other categories also exhibit notable recovery with the longer temporal input. For example, *falling* (A43) shows confusion with *staggering* (A42) under the 16-frame configuration, whereas the 32-frame matrix presents a substantially stronger diagonal response. Likewise, *make a phone call/answer phone* (A28) is more frequently confused with *playing with phone/tablet* (A29) at 16 frames, while their separation becomes clearer with 32 frames. In these cases, the additional observations provide a longer description of the action trajectory while preserving the same underlying CD-Former architecture.

When the temporal input is increased to 32 frames, CD-Former reaches $89.28 \pm 0.10\%$ X-Sub and $92.96 \pm 0.12\%$ X-Set accuracy at an analytical cost of 23.87 GFLOPs. As illustrated in Fig. 3(b), the improvement is reflected not only in a stronger overall diagonal concentration, but also in the reduction of several prominent class-specific confusions observed at 16 frames. The behavior of A20/A21 and A87/A88 is particularly consistent with the importance of temporal progression for distinguishing actions that traverse related skeletal configurations in different directions.

These observations suggest that the additional temporal context available to the 32-frame configuration is especially relevant for actions whose discrimination depends on motion progression, temporal ordering, or the separation of closely related movement patterns. This interpretation is consistent with the aggregate increase in recognition accuracy observed between the two temporal endpoints. However, the comparison remains descriptive: the confusion matrices do not establish a causal contribution of temporal length to any individual action category, nor do they constitute a class-wise statistical significance analysis.

The comparison also complements the accuracy-computation landscape presented in Fig. 2. The 16-frame configuration offers the lowest analytical operation count among the evaluated CD-Former settings, whereas the 32-frame configuration provides the strongest aggregate recognition performance and reduces several of the class-specific confusions visible under the shorter input. Fig. 3 therefore exposes the class-wise dimension of the same trade-off between temporal context, recognition performance, and computational demand observed in the aggregate results.

4.5 Runtime and Edge Deployment Analysis

Beyond classification accuracy, the study characterizes model-only inference on an embedded NVIDIA Jetson TX2. The measurements report single-sample throughput, model inference latency, and GPU memory consumption after the skeleton tensor is available.

Table 4 reports the embedded measurements for the configurations with 16, 24, and 32 frames. All Jetson TX2 values correspond to single-sample inference without batch accumulation. The measured latency covers the recognition model after skeleton data are available and formatted as input tensors; it excludes RGB-to-skeleton pose estimation, disk I/O, network communication, and external sensor acquisition. Because the retained experiment records contain aggregate timing values rather than repeat-level timing samples, runtime standard deviations are not reported.

Table 4: Embedded model-only inference measurements for CD-Former on the NVIDIA Jetson TX2. Recognition accuracy for the same temporal configurations is reported in Table 3. Runtime and memory entries are measured aggregate values; CD-Former GFLOPs are analytically calculated.

Hardware	Frames	Analytical GFLOPs	Throughput (samples/s)	Latency (ms/sample)	VRAM (GB)
Jetson TX2	16	10.47	17.1	59	0.11
Jetson TX2	24	16.80	10.4	96	0.18
Jetson TX2	32	23.87	7.1	140	0.25

Note: Timing values correspond to measured single-sample, model-only inference after the skeleton tensor is available. Throughput is reported as processed samples per second rather than video frames per second. The retained runtime records contain aggregate timing values rather than repeat-level timing samples; therefore, runtime standard deviations cannot be reconstructed reliably and are not reported. The analytical GFLOP counts use a two-FLOPs-per-MAC convention and include the principal input-projection, self-attention, feed-forward, and classification matrix operations.

At 16 frames, the measured model-only throughput is 17.1 samples/s with 59 ms latency and 0.11 GB of VRAM usage. At 32 frames, the corresponding measurements are 7.1 samples/s, 140 ms, and 0.25 GB of VRAM.

From a computational standpoint, the quadratic attention complexity scales with the number of tokens, i.e., $\mathcal{O}((TJ)^2)$, where T is the number of frames and J the number of joints. Restricting the temporal windows to 16–32 frames reduces the number of token interactions relative to longer temporal inputs; the corresponding recognition results are reported separately for each operating point. The corresponding analytical costs of 10.47, 16.80, and 23.87 GFLOPs provide explicit computational operating points for the configurations with 16, 24, and 32 frames.

Arithmetic operation count and measured runtime characterize different aspects of computational behavior. GFLOPs quantify the principal mathematical operations of the architecture, whereas practical latency additionally depends on hardware utilization, memory access, kernel implementation, batch size, and numerical precision. Accordingly, the analytical GFLOP values and measured latency results are reported separately rather than being treated as interchangeable indicators of deployment performance.

Recent work has examined Transformer deployment for human activity recognition on tiny microcontroller-class devices and reported limitations under those resource constraints [50]. The measurements in this study address a different hardware setting: they characterize model-only CD-Former inference with short temporal windows ($T \in \{16, 24, 32\}$ frames) on an NVIDIA Jetson TX2.

4.6 Qualitative Analysis: Attention Visualization

Fig. 4 presents an illustrative visualization of the spatiotemporal attention patterns of the proposed model across three representative action categories from NTU RGB+D 120. The attention maps are extracted from the final Transformer encoder layer and averaged across all attention heads. For each example, we visualize the temporal attention heatmap by averaging attention responses over joints, and the joint attention map by averaging attention responses over time. The three examples were selected from correctly classified test samples to represent diverse motion patterns, including upper-body motion, object-directed arm motion, and lower-body motion, respectively. The visualization is intended as qualitative evidence of the learned attention behavior and is not used for quantitative evaluation.

The temporal attention heatmaps show higher attention weights at different motion phases in the selected examples. For *Sneeze*, higher responses occur at frames 24–32, where visible head and arm movement occurs. For *Shoot a gun*, higher responses occur during the arm-extension phase at frames 24–32. For *Side kick*, higher responses occur around the leg-extension phase, with lower responses in the displayed preparatory poses.

The joint attention maps provide a qualitative indication of the anatomical regions emphasized

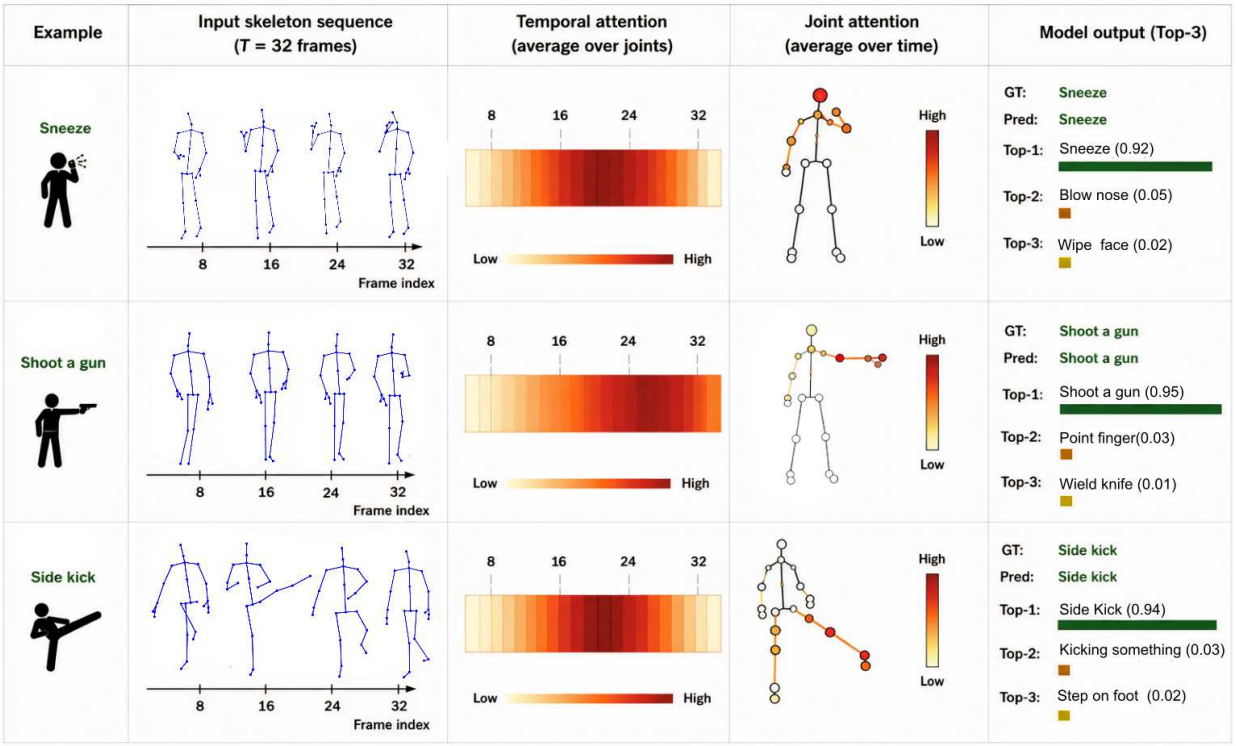


Figure 4: Illustrative visualization of the spatiotemporal attention patterns of the proposed model ($T = 32$ frames) across three correctly classified test samples from NTU RGB+D 120. Attention maps are extracted from the final Transformer encoder layer and averaged across all attention heads. *Left*: input skeleton sequences at frames 8, 16, 24, and 32. *Center*: temporal attention heatmaps averaged over joints (warmer colors indicate higher attention weight). *Right*: joint attention maps averaged over time, overlaid on the skeleton structure. *Far right*: representative Top-3 predictions with confidence scores. The visualization is descriptive: it shows where higher attention weights occur for these examples and is not interpreted as evidence of causal feature importance.

by the model. Upper-body joints, including shoulders, elbows, and wrists, receive comparatively strong responses for *Sneeze* and *Shoot a gun*, whereas lower-body joints, including the hip, knee, and ankle, receive stronger responses for *Side kick*. These observations are illustrative and should not be interpreted as quantitative evidence of causal joint importance.

5 Conclusion

CD-Former illustrates the potential of a comparatively direct pure Transformer formulation for skeleton-based action recognition under short temporal windows. Within the reported evaluation setting, contextual joint-frame representations and standard multi-head self-attention support spatiotemporal modeling without auxiliary graph, convolutional, recurrent, or multibranch processing modules. Across the evaluated configurations, longer temporal inputs are associated with higher recognition accuracy; specifically, the gain from 24 to 32 frames is smaller than the gain from 16 to 24 frames. Measurements on the NVIDIA Jetson TX2 further characterize model-only inference on embedded hardware, providing latency and memory observations relevant to edge deployment.

At the same time, the current formulation leaves substantial room for reducing model size and computational demand. Shorter temporal inputs decrease the number of token interactions, but the feed-forward layers still account for a considerable share of the parameters, while global self-attention retains quadratic complexity with respect to token count. The lower parameter

counts and GFLOP budgets reported by recent specialized and hybrid architectures provide useful reference points for further refinement. Rather than motivating additional architectural complexity, these results encourage the development of more economical pure Transformer representations that preserve the recognition capability observed in this study.

The interpretation of these findings is also bounded by the reported checkpoint-selection procedure and by frame-wise normalization, which suppresses global translation cues. More broadly, the results reinforce the importance of treating edge deployment as a design objective in skeleton-based action recognition rather than only as a post-training evaluation step. Future work will therefore investigate more compact feed-forward layers, reduced token redundancy, and other mechanisms for lowering computation while retaining the direct pure Transformer formulation. The broader objective is to narrow the remaining resource gap while preserving effective spatiotemporal modeling for embedded skeleton-based action recognition.

Acknowledgements

Portions of the research in this paper used the NTU RGB+D and NTU RGB+D 120 Action Recognition Datasets made available by the ROSE Lab at Nanyang Technological University, Singapore.

Data availability

The experiments in this study use the preprocessed 3D skeleton annotations for NTU RGB+D and NTU RGB+D 120 distributed through the PySKL project [44]. The exact annotation files used for these benchmarks are publicly available from the PySKL data resources at `https://download.openmmlab.com/mmaction/pyskl/data/nturgbd/ntu60_3danno.pkl` and `https://download.openmmlab.com/mmaction/pyskl/data/nturgbd/ntu120_3danno.pkl`. The corresponding preprocessing documentation is available at `https://github.com/kennymckormick/pyskl/tree/main/tools/data`. These annotations are derived from the original NTU RGB+D and NTU RGB+D 120 datasets provided by the ROSE Lab at Nanyang Technological University. Information on the original datasets and their access conditions is available at `https://rose1.ntu.edu.sg/dataset/actionRecognition/`. Access to the original datasets requires account registration, submission of a dataset access request, acceptance of the corresponding Release Agreement, and approval by the dataset provider. No new dataset was generated as part of this study, and the original datasets are not redistributed by the authors.

Code availability

The implementation of CD-Former is available at `https://github.com/rombaldivia/CD_FORMER`.

Author contributions

Romel Antonio Baldivia Calderón de la Barca: Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Visualization, Writing–original draft, Writing–review and editing, Project administration. Daniel Ramiro Humberto Calderón Rocabado: Conceptualization, Methodology, Formal analysis, Validation, Supervision, Project administration, Writing–review and editing. Gabriel Antonio Rojas Silva: Methodology, Formal analysis, Validation, Supervision, Writing–review and editing. Miguel Angel Clavijo Quispe: Formal analysis, Validation, Supervision, Writing–review and editing. Xavier Alexis Murillo Sánchez: Methodology, Software, Formal analysis, Writing–review and editing. Guillermo Enrique Manning Soria Galvarro: Software, Investigation, Validation. Carlos Helsner Menacho Guerra: Software, Investigation, Validation. Diego Cesar Calderón Iturricha: Formal analysis, Validation, Supervision, Writing–review and editing.

Competing interests

The authors declare no competing interests.

Funding

The authors received no specific funding for this work.

References

[1] Changyu Liu, Hao Huang, Guogang Huang, Chunyin Wu, and Yingqi Liang. Abnormal action detection based on parameter-efficient transfer learning in laboratory scenarios. *Computers, Materials & Continua*, 80(3):4219–4242, 2024. doi: 10.32604/cmc.2024.053625.

[2] Julian Koch, Lukas Büsch, Martin Gomse, and Thorsten Schüppstuhl. A methods-time-measurement based approach to enable action recognition for multi-variant assembly in human-robot collaboration. *Procedia CIRP*, 106:233–238, 2022. doi: 10.1016/j.procir.2022.02.184.

[3] Bruno Degardin, Vasco Lopes, and Hugo Proença. ATOM: Self-supervised human action recognition using atomic motion representation learning. *Image and Vision Computing*, 137: 104750, 2023. doi: 10.1016/j.imavis.2023.104750.

[4] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208–209:103219, 2021. doi: 10.1016/j.cviu.2021.103219.

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. doi: 10.1109/TPAMI.2019.2916873.

[6] Ruiqi Sun, Qin Zhang, Chuang Luo, Jiamin Guo, and Hui Chai. Human action recognition using a convolutional neural network based on skeleton heatmaps from two-stage pose estimation. *Biomimetic Intelligence and Robotics*, 2(3):100062, 2022. doi: 10.1016/j.birob.2022.100062.

[7] Yong Du, Yun Fu, and Liang Wang. Skeleton-based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583, 2015. doi: 10.1109/ACPR.2015.7486569.

[8] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3D action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4570–4579, 2017. doi: 10.1109/CVPR.2017.486.

[9] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. doi: 10.1109/CVPR.2015.7298714.

[10] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 7444–7452, 2018. doi: 10.1609/aaai.v32i1.12328.

[11] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12018–12027, 2019. doi: 10.1109/CVPR.2019.01230.

[12] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13339–13348, 2021. doi: 10.1109/ICCV48922.2021.01311.

[13] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. STAR-Transformer: A spatio-temporal cross-attention transformer for human action recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3319–3328, 2023. doi: 10.1109/WACV56688.2023.00333.

[14] Junuk Cha, Muhammad Saqlain, Donguk Kim, Seungeun Lee, Seongyeong Lee, and Seungryul Baek. Learning 3D skeletal representation from transformer for action recognition. *IEEE Access*, 10:67541–67550, 2022. doi: 10.1109/ACCESS.2022.3185058.

[15] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537:164–186, 2023. doi: 10.1016/j.neucom.2023.03.001.

[16] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3D skeleton-based action recognition using learning method. *Cyborg and Bionic Systems*, 5:0100, 2024. doi: 10.34133/cbsystems.0100.

[17] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 143–152, 2020. doi: 10.1109/CVPR42600.2020.00022.

[18] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. doi: 10.1109/TIP.2020.3028207.

[19] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. InfoGCN: Representation learning for human skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022. doi: 10.1109/CVPR52688.2022.01955.

[20] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10410–10419, 2023. doi: 10.110 9/ICCV51070.2023.00958.

[21] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Computer Vision – ACCV 2020*, pages 38–53, 2020. doi: 10.1007/978-3-030-69541-5_3.

[22] Yifan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3229–3237, 2021. doi: 10.1145/3474085.3475 473.

[23] Zhimin Gao, Peitao Wang, Pei Lv, Xiaoheng Jiang, Qidong Liu, Pichao Wang, Mingliang Xu, and Wanqing Li. Focal and global spatial–temporal transformer for skeleton-based action recognition. In *Computer Vision – ACCV 2022*, pages 155–171. Springer, 2022. doi: 10.1007/978-3-031-26316-3_10.

[24] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal segments attention for skeleton-based action recognition. *Neurocomputing*, 518:30–38, 2023. doi: 10.1016/j.neucom.2022.10.084.

[25] Wenhan Wu, Ce Zheng, Zihao Yang, Chen Chen, Srijan Das, and Aidong Lu. Frequency guidance matters: Skeletal action recognition by frequency-aware mixed transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4660–4669. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3681009.

[26] Wentian Xin, Qiguang Miao, Yi Liu, Ruyi Liu, Chi-Man Pun, and Cheng Shi. Skeleton mixformer: Multivariate topology representation for skeleton-based action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2211–2220. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3611900.

[27] Tasweer Ahmad, Syed Tahir Hussain Rizvi, and Neel Kanwal. Transforming spatio-temporal self-attention using action embedding for skeleton-based action recognition. *Journal of Visual Communication and Image Representation*, 95:103892, 2023. doi: 10.1016/j.jvcir.2023.103892.

[28] Caojie Xu, Yue Zhou, Aichun Zhu, Zixuan Wang, Yifeng Li, and Fangqiang Hu. ActionFormer: Pose-based action recognition via transformer and CNN. In *2022 China Automation Congress (CAC)*, pages 36–41, 2022. doi: 10.1109/CAC57257.2022.10055176.

[29] Lei Wang and Piotr Koniusz. 3Mformer: Multi-order multi-mode transformer for skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5620–5631, 2023. doi: 10.1109/CVPR52729.2023.00544.

[30] Abhisek Ray, Ayush Raj, and Maheshkumar H. Kolekar. Autoregressive adaptive hypergraph transformer for skeleton-based activity recognition. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9690–9699, 2025. doi: 10.1109/WACV61041.2025.00938.

[31] Jeonghyeok Do and Munchurl Kim. SkateFormer: Skeletal-temporal transformer for human action recognition. In *Computer Vision – ECCV 2024*, volume 15099 of *Lecture Notes in Computer Science*, pages 401–420. Springer, 2025. doi: 10.1007/978-3-031-72940-9_23.

[32] Zijie Zhu, Rendong Ying, Fei Wen, and Peilin Liu. SPAR: An efficient self-attention network using switching partition strategy for skeleton-based action recognition. *Neurocomputing*, 562:126915, 2023. doi: 10.1016/j.neucom.2023.126915.

[33] Zhifu Zhao, Ziwei Chen, Jianan Li, Xuemei Xie, Kai Chen, Xiaotian Wang, and Guangming Shi. STDM-Transformer: Space-time dual multi-scale transformer network for skeleton-based action recognition. *Neurocomputing*, 563:126903, 2024. doi: 10.1016/j.neucom.2023.126903.

[34] Ruyi Liu, Yu Chen, Feiyu Gai, Yi Liu, Qiguang Miao, and Shuai Wu. Local and global spatial–temporal transformer for skeleton-based action recognition. *Neurocomputing*, 634:129820, 2025. doi: 10.1016/j.neucom.2025.129820.

[35] Yiliang Liu, Guangkuo Gao, Chen Pang, and Lei Lyu. Multi-view transformer with hierarchical attention for action recognition. *Neurocomputing*, 700:134402, 2026. doi: 10.1016/j.neucom.2026.134402.

[36] Junyi Wang, Ziao Li, Bangli Liu, Haibin Cai, Mohamad Saada, and Qinggang Meng. High-performance inference graph convolutional networks for skeleton-based action recognition. *Neurocomputing*, 653:131078, 2025. doi: 10.1016/j.neucom.2025.131078.

[37] Yi Liu, Ruyi Liu, Yuzhi Hu, Mengyao Wu, Wentian Xin, Qiguang Miao, Shuai Wu, and Long Li. A systematic review of skeleton-based action recognition: Methods, challenges, and future directions. *IEEE Transactions on Neural Networks and Learning Systems*, 37(5):2046–2065, 2026. doi: 10.1109/TNNLS.2025.3632689.

[38] Xiaofei Qin, Rui Cai, Jiabin Yu, Changxiang He, and Xuedian Zhang. An efficient self-attention network for skeleton-based action recognition. *Scientific Reports*, 12(1):4111, 2022. doi: 10.1038/s41598-022-08157-5.

[39] Yimeng Zhang, Yang Yang, and Xuehao Gao. Lightweight graph convolutional network for efficient skeleton-based action recognition. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10651467.

[40] Qifan Sun, Zenan Fu, Lei Zhang, Guangjie Chen, Wenbo Huang, Hao Wu, and Aiguo Song. ActiFormer: Sign-aware linear attention for sensor-based human activity recognition. *IEEE Internet of Things Journal*, 13(2):3482–3499, 2026. doi: 10.1109/JIOT.2025.3633705.

[41] Qifan Sun, Zechen Li, Kun Wang, Zenan Fu, Guangjie Chen, Lei Zhang, Hao Wu, and Aiguo Song. TSA-Former: Linear transformer with taylor series attention for sensor-based human activity recognition. *IEEE Internet of Things Journal*, 13(15):33538–33553, 2026. doi: 10.1109/JIOT.2026.3694938.

[42] Xin Liu, Dongzhou Cheng, Zenan Fu, Lei Zhang, Hao Wu, and Aiguo Song. Sensor-prompt tuning: Aligning time series foundational models with motion sensors for few-shot activity recognition. *IEEE Transactions on Mobile Computing*, 2026. doi: 10.1109/TMC.2026.3664424.

[43] Zenan Fu, Dongzhou Cheng, Lei Zhang, Wenbo Huang, Zhenghao Chen, and Hao Wu. DeepSenseMoE: Harnessing power of time series foundation models for few-shot human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 292–299, 2026. doi: 10.1609/aaai.v40i1.36990.

[44] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. PYSKL: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3548546.

[45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. doi: 10.1109/CVPR.2016.115.

[46] Zhenhua Li, Fanjia Li, and Gang Hua. Dynamic graph attention network for skeleton-based action recognition. *Applied Sciences*, 15(9):4929, 2025. doi: 10.3390/app15094929.

[47] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. BlockGCN: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2058, 2024. doi: 10.1109/CVPR52733.2024.00200.

[48] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978, 2022. doi: 10.1109/CVPR52688.2022.00298.

[49] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1915–1925, 2021. doi: 10.1109/TCSVT.2020.3015051.

[50] Emanuele Lattanzi, Lorenzo Calisti, and Chiara Contoli. Are transformers a useful tool for tiny devices in human activity recognition? In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pages 339–344. Association for Computing Machinery, 2024. doi: 10.1145/3704137.3704171.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- metricsCDFORMER.zip