computation

MDPI

Article

DPCA-GCN: Dual-Path Cross-Attention Graph Convolutional Networks for Skeleton-Based Action Recognition

Khadija Lasri *, Khalid El Fazazy, Adnane Mohamed Mahraz, Hamid Tairi and Jamal Riffi

I3A Laboratory, Department of Computer Sciences, Faculty of Sciences, Dhar El Mahraz, Sidi Mohamed ben Abdellah University, Fez 30050, Morocco

* Correspondence: khadija.lasri@usmba.ac.ma

Abstract

Skeleton-based action recognition has achieved remarkable advances with graph convolutional networks (GCNs). However, most existing models process spatial and temporal information within a single coupled stream, which often obscures the distinct patterns of joint configuration and motion dynamics. This paper introduces the Dual-Path Cross-Attention Graph Convolutional Network (DPCA-GCN), an architecture that explicitly separates spatial and temporal modeling into two specialized pathways while maintaining rich bidirectional interaction between them. The spatial branch integrates graph convolution and spatial transformers to capture intra-frame joint relationships, whereas the temporal branch combines temporal convolution and temporal transformers to model inter-frame dependencies. A bidirectional cross-attention mechanism facilitates explicit information exchange between both paths, and an adaptive gating module balances their respective contributions according to the action context. Unlike traditional approaches that process spatial–temporal information sequentially, our dual-path design enables specialized processing while maintaining cross-modal coherence through memory-efficient chunked attention mechanisms. Extensive experiments on the NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that DPCA-GCN achieves competitive joint-only accuracies of 88.72%/94.31% and 82.85%/83.65%, respectively, with exceptional top-5 scores of 96.97%/99.14% and 95.59%/95.96%, while maintaining significantly lower computational complexity compared to multi-modal approaches.

Keywords: skeleton-based action recognition; graph convolutional networks; dual-path architecture; cross-attention mechanisms; spatial-temporal modeling; NTU RGB+D

check for updates

Academic Editor: Demos T. Tsahalis

Received: 2 November 2025
Revised: 4 December 2025
Accepted: 7 December 2025
Published: 15 December 2025

Citation: Lasri, K.; El Fazazy, K.; Mohamed Mahraz, A.; Tairi, H.; Riffi, J. DPCA-GCN: Dual-Path Cross-Attention Graph Convolutional Networks for Skeleton-Based Action Recognition. *Computation* **2025**, *13*, 293. https://doi.org/10.3390/computation13120293

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Human action recognition (HAR) represents a fundamental challenge in computer vision with extensive applications across diverse domains, including intelligent surveillance systems [1], human–computer interaction [2], healthcare monitoring [3], and autonomous robotics [4]. While RGB-based action recognition has demonstrated significant progress, it faces several critical limitations: sensitivity to illumination variations [1], susceptibility to background clutter [2], and performance degradation under occlusions [3]. In contrast, skeleton-based representation offers a more robust and efficient alternative by focusing on essential structural information of human poses, providing invariance to appearance changes and background interference while maintaining computational efficiency [4].

The evolution of skeleton-based action recognition reflects a progression from simple to increasingly sophisticated approaches. Early methods relied on geometric features and

statistical descriptors to capture motion patterns, including joint trajectories [5], relative position encoding [6], and hierarchical body part analysis [7]. As deep learning emerged, researchers explored sequential models like RNNs and frame-based CNNs, but these architectures struggled to fully capture the inherent graph structure of the human skeleton, where joints and bones form a natural hierarchical representation of human motion.

The emergence of graph convolutional networks (GCNs) has revolutionized skeleton-based action recognition by providing a natural framework for processing graph-structured data [8]. The seminal Spatial-Temporal Graph Convolutional Network (ST-GCN) introduced by Yan et al. [9] established the foundation for modern skeleton-based methods by simultaneously modeling spatial joint relationships and temporal dynamics through graph convolution operations. ST-GCN constructs spatial–temporal graphs where nodes represent joints across different time frames, and edges encode both spatial anatomical connections and temporal correspondences. This approach significantly outperformed previous methods by effectively capturing the spatial–temporal dependencies inherent in human actions.

Building upon ST-GCN's foundation, the field has seen rapid advancement through several key innovations. Early improvements focused on enhancing graph topology: Adaptive Graph Convolutional Networks [10] introduced learnable adjacency matrices to capture dynamic joint relationships, while Dynamic GCN [11] proposed adaptive graph construction based on joint correlations. Subsequent works explored multi-stream architectures, with AS-GCN [12] separating actional and structural information, and Shift-GCN [13] introducing efficient shift operations for lightweight processing. Recent advances have pushed performance further through sophisticated feature learning: CTR-GCN [14] employs channel-wise topology refinement, InfoGCN [15] leverages information bottleneck principles for better feature extraction, and MS-GCN [16] introduces multi-scale graph convolutions for hierarchical feature learning. Despite these advances, several fundamental challenges remain unaddressed in current approaches.

Current approaches face three key challenges in skeleton-based action recognition. First, most existing methods process spatial and temporal information through a single pathway, limiting their ability to capture the distinct characteristics of joint configurations and motion patterns [14,15]. This unified processing can blur important modality-specific features that require specialized treatment. Second, as sequence lengths grow, memory requirements become prohibitive due to the quadratic complexity of attention mechanisms [17,18], restricting the model's ability to handle long-term dependencies. Third, the lack of explicit interaction between spatial and temporal features means that valuable cross-modal patterns may be missed, potentially reducing recognition accuracy for complex actions that depend on both pose and motion cues [19].

To address these limitations, we introduce the dual-path cross-attention GCN (DPCA-GCN). Our architecture offers three key innovations: (1) dedicated spatial and temporal pathways, each optimized for its specific domain; (2) memory-efficient bidirectional cross-attention that enables selective information exchange between pathways; and (3) an adaptive gating mechanism that automatically balances spatial and temporal contributions based on action context. Extensive experiments on the NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that DPCA-GCN achieves performance comparable with recent state-of-the-art methods, validating the effectiveness of explicit dual-path modeling and targeted cross-modal fusion for skeleton-based action recognition. The main contributions of this paper are summarized as follows:

- We propose a novel dual-path architecture that explicitly separates spatial and temporal feature extraction through specialized pathways, enabling modality-specific

learning that better captures the distinct characteristics of joint configurations and motion dynamics.

- We introduce a memory-efficient bidirectional cross-attention mechanism that reduces complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L \cdot C)$ while enabling rich information exchange between spatial and temporal pathways at multiple processing stages.
- We design an adaptive gating module that learns action-specific fusion weights, dynamically balancing spatial and temporal contributions according to input characteristics rather than using fixed fusion strategies.
- We conduct extensive experiments and ablation studies on NTU RGB+D 60 and NTU RGB+D 120 datasets, achieving competitive joint-only performance (88.72%/94.31% and 82.85%/83.65%) with significantly lower computational complexity compared to multi-modal approaches, along with exceptional top-5 accuracies (96.97%/99.14% and 95.59%/95.96%).

The remainder of this paper is organized as follows. Section 2 reviews related work in skeleton-based action recognition and graph neural networks. Section 3 details our proposed DPCA-GCN approach. Section 4 presents experimental results and analysis. Finally, Section 5 concludes the paper and discusses future work.

2. Related Work

The field of skeleton-based human action recognition has evolved significantly, particularly in the past five years, through innovations in graph neural networks, attention mechanisms, and efficient processing techniques. This section systematically reviews the key developments that inform our DPCA-GCN approach.

2.1. Graph Convolutional Networks for Skeleton Data

Modern skeleton-based action recognition has been fundamentally shaped by advances in graph convolutional networks, which provide natural frameworks for modeling the structural relationships inherent in human skeletal data. Graph topology enhancement represents a critical advancement in this domain, with several key methodological developments emerging.

Chen et al. [14] introduced CTR-GCN with channel-wise topology refinement, demonstrating that adaptive adjacency matrices significantly improve performance through better graph representation learning. This approach addresses the limitation of fixed graph topologies by enabling networks to learn task-specific spatial relationships between joints. Building upon similar principles, Chi et al. [15] proposed InfoGCN, which applies information bottleneck theory to enhance feature discriminability while reducing redundancy in graph representations. These topology-aware methods established the importance of adaptive graph construction, though they maintained unified processing pathways for spatial and temporal information.

Multi-scale feature learning has emerged as another fundamental technique for capturing hierarchical motion patterns. Kilic et al. [16] developed MS-GCN with multi-scale convolutions to capture features at different granularities, enabling models to understand both fine-grained joint movements and broader action patterns. Liu et al. [20] proposed TSGCNeXt, combining dynamic and static graph convolutions for efficient long-term modeling. Their approach demonstrates the effectiveness of hybrid strategies that leverage both fixed skeletal structure and adaptive relationships.

Recent developments have focused on advanced architectural designs and processing mechanisms. Zhou et al. [21] introduced BlockGCN, which redefines topology awareness through block-wise processing to better capture local and global motion patterns. Myung et al. [22] proposed deformable graph convolutions that dynamically adjust to varying

skeleton structures, showing remarkable adaptability to different body configurations and motion styles. Most recently, Li and Li [23] proposed DF-GCN (2025), a lightweight dynamic dual-stream network that explicitly separates spatial and temporal processing with dynamic graph construction, underscoring the persistent value of dual-stream architectures. Similarly, Chen et al. [24] introduced a two-stream GCN–Transformer hybrid (2025) combining graph convolutions with transformer attention, validating the effectiveness of hybrid strategies for balancing local structural modeling and global context capture. Our own recent work [25] further validates the hybrid approach by demonstrating the efficacy of attention-inflated 3D architectures for capturing complex action dynamics.

Beyond skeleton-specific methods, the broader action recognition literature provides valuable insights. Ramanathan et al. [26] developed a mutually reinforcing motion–pose framework for pose-invariant recognition from RGB videos, demonstrating that explicit modeling of pose and motion as separate but interacting components improves robustness. In our comparative study of I3D and SlowFast networks [27], we observed similar trade-offs between spatial and temporal modeling in video data, reinforcing the need for architectures that can effectively balance these modalities. While their work focuses on RGB-based recognition without explicit skeletons, it validates our core hypothesis that motion and pose (analogous to our temporal and spatial features) benefit from separate processing with explicit interaction mechanisms.

A common characteristic across existing graph-based methods is their reliance on unified or sequential pathways for processing spatial and temporal information. While effective, this architectural choice constrains the ability to capture modality-specific patterns that require specialized treatment. Our DPCA-GCN addresses this limitation through explicit dual-path processing with bidirectional cross-attention fusion, enabling domain-specific optimizations while maintaining cross-modal coherence.

2.2. Attention and Transformer-Based Models

The integration of attention mechanisms into skeleton-based models has progressed from simple hybrid approaches to sophisticated transformer architectures. Early hybrid approaches focused on enhancing GCN capabilities. Plizzari et al. [28] developed a spatial–temporal attention module that selectively emphasizes informative joints and frames. Si et al. [29] proposed AGC-LSTM, augmenting LSTM cells with joint-level attention to capture long-term dependencies in motion sequences.

Pure-attention architectures, primarily Transformer-based models, emerged as powerful alternatives to GCN-based approaches. Plizzari et al. [30] introduced ST-TR, which innovatively tokenizes joint–frame pairs through a Transformer encoder, enabling direct modeling of global spatio-temporal relationships. Shi et al. [31] developed DSTA-Net with factorized attention mechanisms, separately processing spatial and temporal information with decoupled positional encodings. This design significantly reduced computational complexity while maintaining high accuracy. Ahn et al. [32] proposed STAR-Transformer, introducing a sparse neighborhood attention pattern that achieves linear computational complexity without sacrificing performance.

However, transformer-based architectures faces significant limitations for skeleton-based recognition: (1) Loss of structural inductive bias—transformers treat skeleton as unordered token sets, discarding the natural graph topology that GCNs explicitly model. This requires substantially more training data to learn spatial relationships that GCNs encode architecturally. (2) Computational inefficiency—despite optimization efforts, attention mechanisms remain memory-intensive ($\mathcal{O}(L^2)$ or $\mathcal{O}(L \cdot \sqrt{L})$ for efficient variants), limiting their applicability to long sequences or resource-constrained scenarios. (3) Difficulty in incorporating domain knowledge—skeletal topology (bone connections, body part hierar-

chies) cannot be naturally integrated into pure-attention frameworks, unlike GCNs where adjacency matrices encode such priors.

Recent transformer-based innovations have focused on efficiency and scalability. Liu et al. [20] introduced TSGCNeXt in 2023, combining dynamic and static graph convolutions within a transformer framework for efficient long-term modeling. Their approach particularly excels at capturing both consistent skeletal structure and varying temporal dynamics. The influence of video-centric transformers has been significant, with Bertasius et al. [33] proposing TimeSformer, Arnab et al. [34] developing ViViT, and Fan et al. [35] introducing MViT. These works have inspired various efficient variants using techniques like pooled attention and chunk-based processing, addressing the computational challenges of processing long sequences.

Our approach leverages the complementary strengths of both paradigms: GCNs provide efficient structural modeling with built-in skeletal priors, while transformers enable flexible global context capture. This hybrid strategy has proven effective across recent methods [20,24], suggesting that combining structural inductive biases with attention mechanisms offers superior efficiency–accuracy trade-offs compared to transformer-based approaches.

2.3. Multi-Stream Spatio-Temporal Frameworks

The concept of processing complementary motion aspects through separate streams has proven highly effective. Feichtenhofer et al. [36] pioneered this approach with SlowFast networks for RGB-based recognition, processing slow semantic features and fast motion patterns in parallel pathways. In the skeleton domain, Liu et al. [37] proposed Disentangled-ST, featuring independent spatial and temporal encoders whose outputs are strategically fused. This separation allows each encoder to specialize in its respective domain while maintaining information exchange through careful fusion strategies.

Recent work has explored increasingly sophisticated multi-stream architectures. Duan et al. [38] developed DG-STGCN, employing dynamic graphs in the spatial stream and grouped convolutions in the temporal stream, achieving state-of-the-art performance through specialized processing pathways. Wang et al. [39] proposed SpSt-GCN, introducing novel information exchange mechanisms between spatial and structural streams that enhance the model's ability to capture complex motion patterns.

The latest advancements focus on optimal stream interaction and fusion. Zhou et al. [21] introduced BlockGCN in 2024, redefining topology awareness through block-wise graph processing that better captures local and global motion patterns. Their approach demonstrates superior performance on actions requiring understanding of both fine-grained movements and overall body coordination.

3. Materials and Methods

3.1. Overview and Architectural Rationale

Most Spatial–Temporal GCN models process joint configuration (space) and motion dynamics (time) in a single, tightly coupled stream, making it difficult to learn modality-specific patterns. DPCA-GCN addresses this limitation with two dedicated pathways—spatial and temporal—that communicate through an efficient bidirectional cross-attention mechanism and are mixed by an adaptive gate.

Our architectural choices directly address the three identified limitations:

(1) Dual-Path Design for Modality-Specific Learning: Spatial and temporal features have fundamentally different characteristics—spatial features capture joint relationships within frames (graph-structured), while temporal features capture motion dynamics across frames (sequence-structured). Processing these jointly in a single pathway forces a compromised representation. Our dual-path design allows (a) the spatial pathway to use graph

convolution optimized for skeletal topology, and (b) the temporal pathway to use 1D convolution optimized for sequential dependencies. This separation follows the multi-stream paradigm proven effective in video understanding [36], where specialized processing of different modalities yields superior performance compared to unified architectures.

(2) Chunked Attention for Memory Efficiency: Standard attention mechanisms require $\mathcal{O}(L^2)$ memory for sequence length L, making them impractical for skeleton sequences where $L = T \times V$ (frames $\times$ joints). For typical settings ($T = 64, V = 25$, resulting in $L = 1600$ tokens per sequence), the quadratic memory dependency of full attention poses a significant computational barrier. Our chunked mechanism addresses this by dividing queries into chunks while maintaining full key–value access, thereby reducing the memory complexity to $\mathcal{O}(L \cdot C)$ where C is chunk size. This innovation enables the processing of long sequences without sacrificing global context, unlike simpler sliding window approaches that inherently lose long-range dependencies.

(3) Bidirectional Cross-Attention for Cross-Modal Interaction: Simply concatenating spatial and temporal features (early fusion) or averaging their predictions (late fusion) fails to capture their intricate dependencies. For example, recognizing "drinking" requires understanding both the hand trajectory (temporal) and hand–mouth spatial proximity (spatial) simultaneously. Bidirectional attention enables each pathway to query the other: spatial features can attend to relevant temporal patterns and vice versa, creating rich cross-modal representations that neither pathway alone could produce.

Figure 1 shows the complete pipeline: an input skeleton sequence is normalized, fed to both paths, fused at each of the L stacked DPCA blocks, and finally aggregated by global average pooling and a linear classifier.

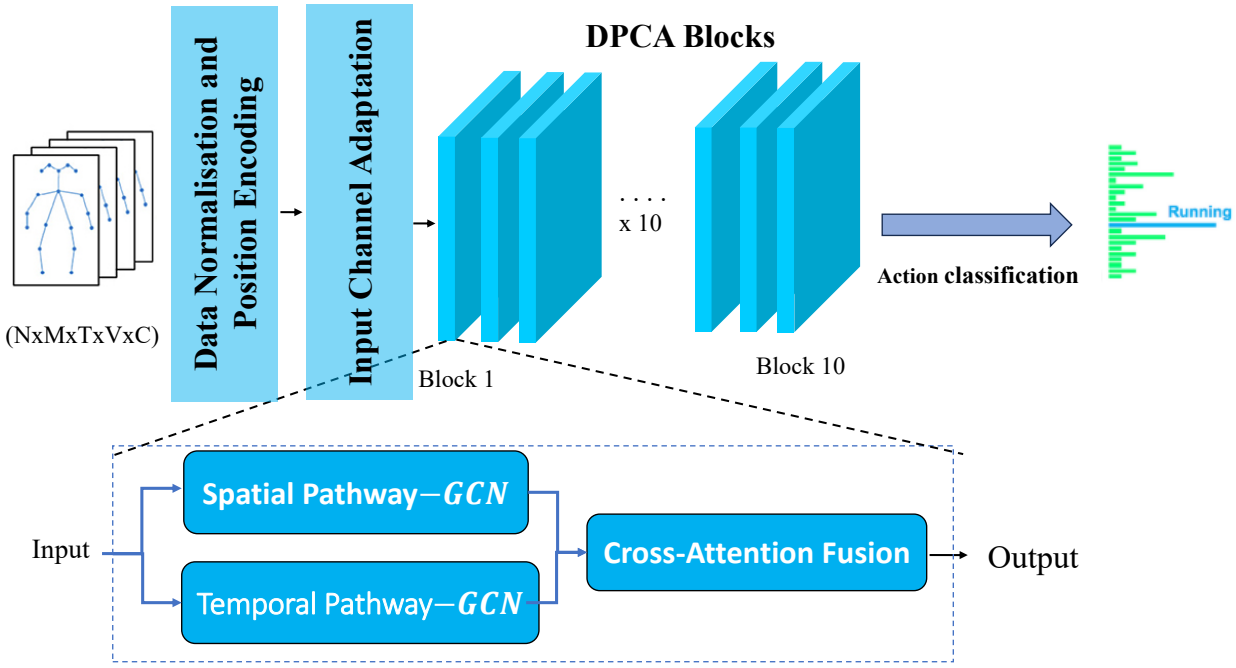


Figure 1. High-level DPCA-GCN pipeline (macro-architecture). This figure illustrates the complete system, where an input skeleton sequence is processed through multiple stacked DPCA blocks. Each block contributes to progressive feature extraction and fusion, leading to a final classification. Figure 2 provides a detailed view of a single DPCA block's internal structure.

3.2. Dual-Path Feature Extraction

Given features $\mathbf{X}_s^{(l)} \in \mathbb{R}^{N\times C\times T\times V}$ at block l, the spatial pathway performs graph convolution to aggregate neighboring joints:

$$\mathbf{H}_s^{(l)} = \sigma(\mathbf{A}_s \mathbf{X}_s^{(l)} \mathbf{W}_s^{(l)} + \mathbf{b}_s^{(l)}) \tag{1}$$

where $\mathbf{A}_s$ is a learnable adjacency matrix and σ denotes ReLU activation. The features are then processed by a spatial transformer:

$$\mathbf{Z}_s^{(l)} = \text{Transformer}(\mathbf{H}_s^{(l)}) \tag{2}$$

Similarly, the temporal pathway applies 1-D convolution along the time axis, followed by a temporal transformer:

$$\mathbf{H}_t^{(l)} = \sigma(\mathbf{W}_t^{(l)} * \mathbf{X}_t^{(l)} + \mathbf{b}_t^{(l)}) \tag{3}$$

$$\mathbf{Z}_t^{(l)} = \text{Transformer}(\mathbf{H}_t^{(l)}) \tag{4}$$

where $*$ denotes 1-D temporal convolution.

3.3. Transformer Cell (Shared)

Both spatial and temporal transformers adopt the pre-LN design of [40]:

$$\mathbf{Y} = \text{LN}(\mathbf{X}), \tag{5}$$

$$\mathbf{U} = \mathbf{Y} + \text{MHA}(\mathbf{Y}), \tag{6}$$

$$\mathbf{Z} = \mathbf{U} + \text{FFN}\big(\text{LN}(\mathbf{U})\big), \tag{7}$$

where LN is LayerNorm, MHA is h-head self-attention, and FFN is a two-layer perceptron with expansion ratio 4. This cell is parameter-shared across time in the spatial path and across joints in the temporal path.

3.4. Bidirectional Chunked Cross-Attention

Figure 2 details one DPCA block. Linear projections produce $Q_s, K_s, V_s = \mathbf{Z}_s^{(l)} \mathbf{W}_{QKV}^{s}$ and $Q_t, K_t, V_t = \mathbf{Z}_t^{(l)} \mathbf{W}_{QKV}^{t}$. Queries are partitioned into non-overlapping chunks of length C:

$$\mathbf{A}_{s\to t} = \text{ChunkAttn}(Q_s, K_t, V_t), \tag{8}$$

$$\mathbf{A}_{t\to s} = \text{ChunkAttn}(Q_t, K_s, V_s), \tag{9}$$

where `ChunkAttn` computes scaled dot-product attention *inside* each chunk, reducing the memory footprint from $\mathcal{O}(L^2)$ to $\mathcal{O}(L\,C)$ while still letting every query view the entire key/value set.

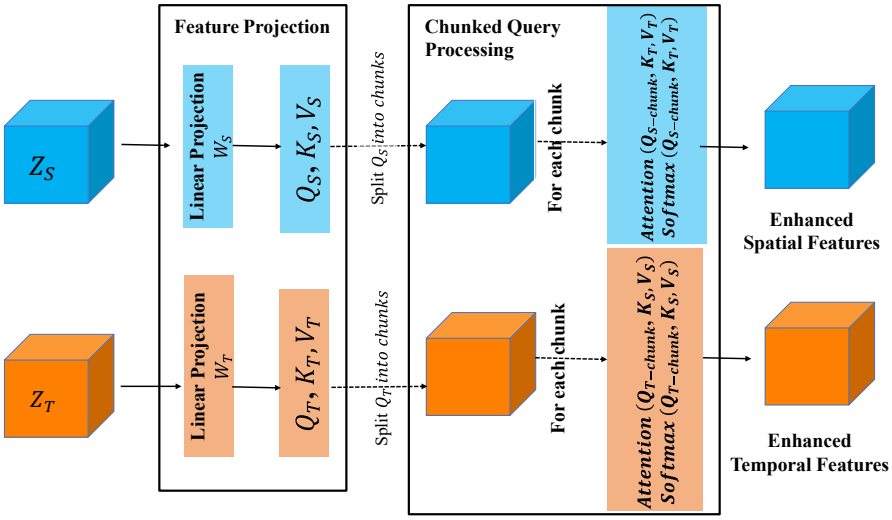


Figure 2. Bidirectional chunked cross-attention mechanism (micro-architecture of a single DPCA block). This figure details the internal components and information flow within one DPCA block, as shown in the macro-architecture (Figure 1). It highlights the feature projection, chunked query processing, and cross-modal information exchange between spatial and temporal pathways.

3.5. Adaptive Gate

The adaptive gate determines the optimal combination of spatial and temporal features based on action context. This module addresses a key limitation of existing multi-stream methods: their inability to dynamically adjust the importance of spatial vs. temporal information according to action characteristics.

Motivation: Different actions exhibit varying spatial–temporal characteristics. For instance, "reading" or "writing" are relatively static with high spatial discriminability (hand-book configuration) but limited temporal dynamics. Conversely, "walking" or "running" show consistent spatial patterns but strong temporal signatures. A fixed fusion strategy cannot optimally handle this diversity.

Implementation: To derive the context-aware fusion weights, the enhanced spatial and temporal feature tensors, $\tilde{\mathbf{X}}_s = \mathbf{Z}_s^{(l)} + \mathbf{A}_{t\to s}$ and $\tilde{\mathbf{X}}_t = \mathbf{Z}_t^{(l)} + \mathbf{A}_{s\to t}$ (where residual connections ensure pathway-specific information is preserved after cross-attention), are first concatenated. This combined representation is then subjected to global average pooling to extract a compact global context vector $\mathbf{g}$:

$$\mathbf{g} = \mathrm{GAP}([\tilde{\mathbf{X}}_s; \tilde{\mathbf{X}}_t]) \in \mathbb{R}^{N\times 2C\times 1\times 1} \tag{10}$$

A two-layer MLP with reduction ratio $r = 16$ (inspired by SENet [14] but adapted for dual-stream fusion) produces channel-wise attention weights:

$$[\alpha, \beta] = \mathrm{softmax}(\mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \mathbf{g})) \tag{11}$$

where $\alpha, \beta \in \mathbb{R}^{N\times C\times 1\times 1}$ and $\alpha + \beta = 1$ (enforced by softmax). The final fusion is computed as

$$\mathbf{X}_{\text{fused}}^{(l)} = \alpha \odot \tilde{\mathbf{X}}_s + \beta \odot \tilde{\mathbf{X}}_t \tag{12}$$

The MLP learns to map input-dependent global statistics $\mathbf{g}$ to fusion weights. For spatially discriminative actions (e.g., "pointing"), the network learns $\alpha > \beta$, emphasizing spatial features. For temporally discriminative actions (e.g., "jumping"), it learns $\alpha < \beta$. This adaptation happens per sample and per layer, allowing hierarchical refinement of spatial–temporal balance as features become more abstract in deeper layers.

While adaptive fusion has been explored in multi-modal learning, our formulation uniquely operates on *cross-attention-enhanced* features rather than raw pathway outputs. This

enables the gate to reason about cross-modal interactions ($\mathbf{A}_{t\to s}$, $\mathbf{A}_{s\to t}$) when determining fusion weights, creating a feedback loop where attention patterns inform fusion decisions.

3.6. Positional Encoding and Network Depth

To keep joint order and frame chronology, we inject learnable spatial and temporal embeddings at the input:

$$\mathbf{X}_0 = \mathbf{X} + \mathbf{PE}_{\text{spatial}} + \mathbf{PE}_{\text{temporal}}.$$

We stack $L = 10$ DPCA blocks; Blocks 5 and 8 double channel width and down-sample time by 2, enlarging the receptive field.

The complete forward pass of DPCA-GCN is summarized in Algorithm 1. Following recent efficient implementations [17,41], we optimize the processing pipeline through careful data normalization, parallel pathway processing, and memory-efficient attention computation. The algorithm consists of four main stages:

1. Input Processing : The input skeleton sequence undergoes data normalization and positional encoding injection, similar to [40] but adapted for spatial–temporal data.
2. Dual-Path Processing: Each DPCA block processes spatial and temporal information in parallel pathways, inspired by [36] but extended for skeleton data.
3. Cross-Modal Fusion : The bidirectional chunked attention mechanism enables efficient information exchange while maintaining linear memory complexity [18].
4. Progressive Feature Aggregation : Multi-scale feature learning is achieved through strategic temporal downsampling and channel inflation, following insights from [19].

Algorithm 1 Forward Pass of DPCA-GCN.

Require: Skeleton sequence $\mathbf{X} \in \mathbb{R}^{N\times T\times V\times C}$ ▷ N: batch, T: frames, V: joints, C: channels
Ensure: Logits $\mathbf{y} \in \mathbb{R}^{N\times cls}$ ▷ cls: number of action classes
$\mathbf{X} \leftarrow \text{DataNorm}(\mathbf{X})$ ▷ Input normalization
$\mathbf{X} \leftarrow \mathbf{X} + \mathbf{PE}_{\text{spatial}} + \mathbf{PE}_{\text{temporal}}$ ▷ Add positional encodings
$\mathbf{X}_s^{(0)}, \mathbf{X}_t^{(0)} \leftarrow \mathbf{X}$ ▷ Initialize pathways
for $l = 0$ **to** $L - 1$ **do** ▷ Process L DPCA blocks
  $\mathbf{Z}_s^{(l)} \leftarrow$ Spatial GCN + Transformer ▷ Spatial pathway via Equation (1)
  $\mathbf{Z}_t^{(l)} \leftarrow$ Temporal Conv + Transformer ▷ Temporal pathway via Equation (3)
  Compute bidirectional cross-attention ▷ Cross-modal exchange
  $\mathbf{X}_{\text{fused}}^{(l)} \leftarrow$ Adaptive gate fusion ▷ Adaptive fusion via Equation (12)
  if $l + 1 \in \{5, 8\}$ **then** ▷ Multi-scale processing
    Apply temporal downsampling and channel inflation
  end if
  $\mathbf{X}_s^{(l+1)}, \mathbf{X}_t^{(l+1)} \leftarrow \mathbf{X}_{\text{fused}}^{(l)}$ ▷ Update features
end for
$\mathbf{h} \leftarrow \text{GlobalAvgPool}(\mathbf{X}_{\text{fused}}^{(L)})$ ▷ Global pooling
$\mathbf{y} \leftarrow \text{LinearClassifier}(\mathbf{h})$ ▷ Classification

4. Results

4.1. Datasets

We evaluate our approach on two widely recognized action recognition datasets: NTU RGB+D 60 [42] and NTU RGB+D 120 [43]. Both datasets are captured using Microsoft Kinect V2 sensors, providing 3D skeletal data with 25 body joint coordinates captured at 30 fps from three camera viewpoints.

4.1.1. NTU RGB+D 60

NTU RGB+D 60 contains 56,880 skeleton sequences across 60 distinct action categories, encompassing 40 daily activities, 9 healthcare-related actions, and 11 interactive behaviors. Two standard evaluation protocols are used:

- Cross-subject (X-Sub): Training set contains 40,320 sequences from 20 subjects (IDs: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38). Testing set contains 16,560 sequences from the remaining 20 subjects. No validation set is explicitly defined; we use the official test set for final evaluation.
- Cross-view (X-View): Training set uses 37,920 sequences from camera views 2 and 3. Testing set uses 18,960 sequences from camera view 1 (frontal view). Following standard practice [9,14], we perform model selection using validation accuracy on a held-out subset of training data (10% randomly sampled), then report final results on the official test set.

4.1.2. NTU RGB+D 120

NTU RGB+D 120 extends the dataset to 120 action classes with 114,480 total sequences. Two evaluation protocols are employed:

- Cross-subject (X-Sub): Training set contains 63,026 sequences from 53 subjects (IDs 1–53 excluding specific held-out subjects). Testing set contains 50,922 sequences from 53 different subjects (IDs 54–106). Validation is performed on 10% of training data following standard practice.
- Cross-setup (X-Setup): Training set uses 54,471 sequences from even-numbered setup IDs (16 setups with different camera configurations and environmental conditions). Testing set uses 59,477 sequences from odd-numbered setup IDs (16 setups). This protocol evaluates generalization across different recording conditions.

Training–Validation–Testing Split: Following established protocols [9,14,15], we use the official train–test splits defined by the dataset creators. For hyperparameter tuning and early stopping, we sample 10% of official training sequences as a validation set (stratified by action class to maintain class balance). Final reported accuracies are computed on the official test sets without any validation data leakage.

Figure 3 illustrates representative action samples from the NTU RGB+D 60 and NTU RGB+D 120 datasets, highlighting the diversity of actions and recording conditions. In addition, Figure 4 shows the configuration and indexing of the 25 body joints used to construct the skeleton representations in both datasets.



Figure 3. Sample action frames from NTU RGB+D datasets. (**Left**): Representative sequences from NTU RGB+D 60 demonstrating various daily activities. (**Right**): Samples from NTU RGB+D 120 showcasing expanded action vocabulary with increased environmental diversity.

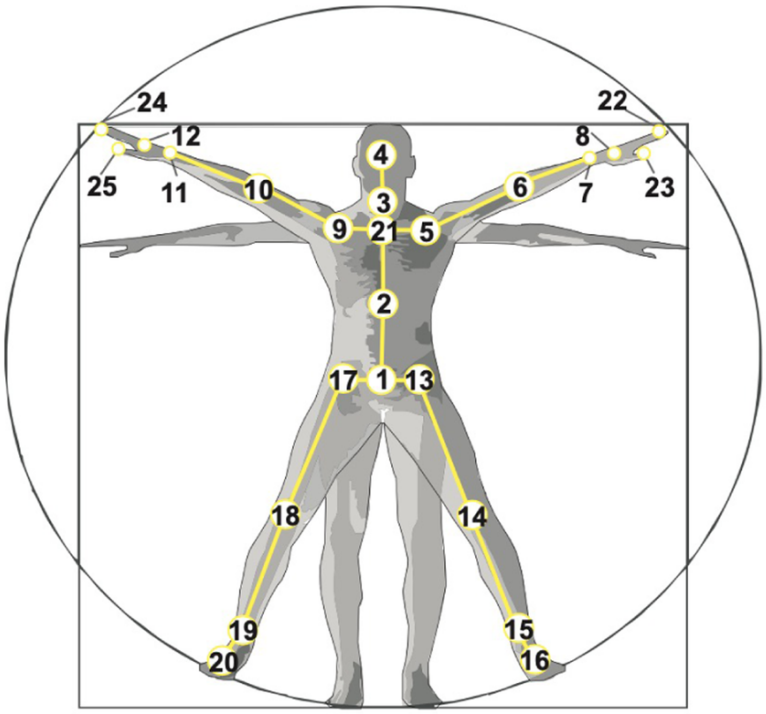


Figure 4. Configuration of 25 body joints used in NTU RGB+D datasets. Joint identification provides comprehensive human pose representation for skeleton-based action analysis. Adapted from [44].

4.2. Implementation Details

Our experiments are implemented using PyTorch 1.10.0 with mixed-precision training for enhanced computational efficiency. The DPCA-GCN model is optimized using AdamW with initial learning rate 0.001 and weight decay 0.05. Training employs Cross-entropy loss with label smoothing (smoothing factor 0.1).

We utilize an effective training batch size of 32 (8 samples per GPU with five-step gradient accumulation) . The learning rate follows cosine annealing, decreasing from the initial rate to 10^{-6} over 45 epochs, including a 12-epoch warmup period. Sequences are uniformly sampled to a fixed length of 64 frames and undergo initial 3D joint coordinate normalization.

Handling Missing and Occluded Joints: Skeleton sequences obtained from depth sensors, such as NTU RGB+D, frequently contain instances of missing or unreliable joint coordinates due to factors like occlusions, sensor limitations, or inaccuracies in pose estimation. Our preprocessing pipeline primarily handles core data preparation steps, including normalizing 3D joint coordinates (`PreNormalize3D` module) and ensuring uniform sequence lengths by sampling frames (`UniformSampleFrames` module).

For any joints that are reported as missing or entirely occluded within the raw input data (e.g., if their coordinates are effectively zeros), our current pipeline implicitly treats these values as zeros after normalization. This means the model processes a sequence where occluded joints effectively contribute zero spatial information. Crucially, our DPCA-GCN architecture, with its transformer-based attention mechanisms, offers an inherent capacity for robustness to such incomplete inputs. These mechanisms can dynamically learn to adaptively weigh or de-emphasize unreliable or zero-valued joint information by adjusting attention scores, effectively learning to infer or ignore patterns from the remaining visible joints.

4.3. Results and Discussion

4.3.1. Performance Evaluation

Table 1 presents comprehensive evaluation results of DPCA-GCN on both datasets. On NTU RGB+D 60, our model achieves 88.72% top-1 accuracy on cross-subject and 94.31% on cross-view protocols, with impressive top-5 accuracies of 96.97% and 99.14%, respectively. On NTU RGB+D 120, the model maintains competitive performance with 82.85% and 83.65% top-1 accuracies, with corresponding top-5 accuracies of 95.59% and 95.96%.

Figure 5 illustrates the training dynamics of DPCA-GCN on the NTU RGB+D 60 X-Sub protocol, including the evolution of the cross-entropy loss and the top-1 and top-5 accuracies across training epochs.

Table 1. Performance of DPCA-GCN on NTU RGB+D datasets, reported from a single comprehensive run. All accuracies are consistently reported to 2 decimal places.

Dataset	Top-1 Accuracy (%)		Top-5 Accuracy (%)	
	X-Sub	X-View/Setup	X-Sub	X-View/Setup
NTU RGB+D 60	88.72	94.31	96.97	99.14
NTU RGB+D 120	82.85	83.65	95.59	95.96

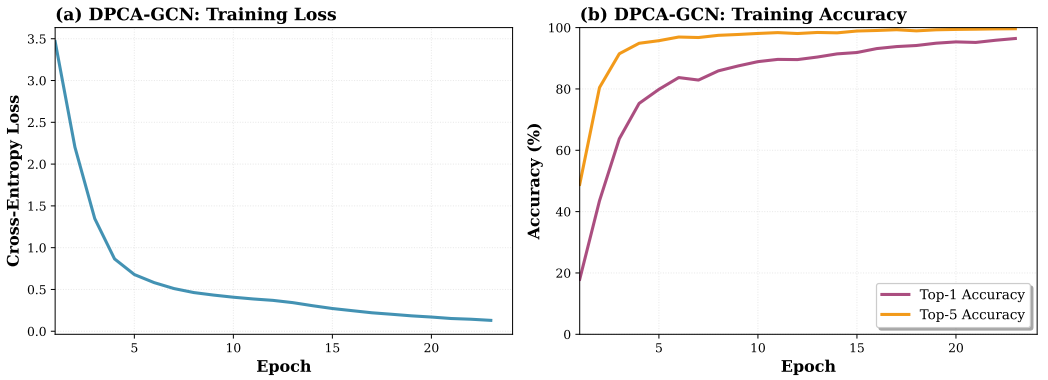


Figure 5. Training dynamics of DPCA-GCN on NTU RGB+D 60 X-Sub protocol. (**a**) Cross-entropy loss decreases smoothly and rapidly from an initial value of 3.5 to below 0.2, demonstrating the robustness and efficiency of our optimization. (**b**) Top-1 and top-5 accuracies show steady improvement. Training was halted at approximately 24 epochs by an early stopping mechanism, ensuring optimal performance without overfitting.

4.3.2. Comparative Analysis

Table 2 compares DPCA-GCN with state-of-the-art methods, organized by modality usage to highlight architectural advantages. Our approach demonstrates competitive performance within the joint-only processing paradigm while introducing significant architectural innovations.

The comparative analysis reveals that DPCA-GCN achieves competitive performance within the joint-only processing paradigm while introducing significant architectural innovations. Our model demonstrates superior performance compared to most existing joint-only methods, particularly excelling in cross-setup evaluation (83.65%) where it surpasses ST-TR (83.4%) and approaches the performance of multi-modal methods with substantially lower computational complexity.

Table 2. Comprehensive comparison of top-1 accuracy (%) with state-of-the-art methods on NTU RGB+D datasets, organized by modality usage to highlight architectural advantages. Baseline methods report accuracies as published in their original papers; our results report 2 decimal places for precision.

Methods	NTU RGB+D 60		NTU RGB+D 120		Modalities
	X-Sub	X-View	X-Sub	X-Set	
Joint-only Methods (Efficient Processing)					
H-RNN [45]	59.1	64.0	-	-	J
VA-LSTM [46]	79.4	87.6	-	-	J
ST-GCN [9]	81.5	88.3	70.7	73.2	J
DPRL+GCNN [47]	83.5	89.8	-	-	J
AS-GCN [12]	86.8	94.2	79.4	81.1	J
GR-GCN [48]	87.5	94.3	-	-	J
RA-GCN [49]	87.3	93.6	-	-	J
Shift-GCN [13]	87.8	**95.1**	-	-	J
SGN [50]	**86.9**	92.8	**79.2**	**81.5**	J
ST-TR [30]	88.70	94.1	82.7	83.4	J
AGC-LSTM [29]	87.5	**93.5**	-	-	J
VA-CNN [51]	**88.7**	**94.3**	-	-	J
DPCA-GCN (Ours)	**88.72**	**94.31**	**82.85**	**83.65**	**J**
Multi-modal Methods (Higher Complexity; Not Directly Comparable)					
CTR-GCN [14]	92.4	96.8	88.9	90.6	J + B + JM + BM
InfoGCN [15]	93.0	97.1	89.8	91.2	J + B + JM + BM
BlockGCN [21]	93.1	97.0	89.9	91.5	J + B + JM + BM

J: Joint; B: Bone; JM: Joint Motion; BM: Bone Motion. **Bold blue**: best in joint-only category; underline: second best. Multi-modal methods use additional input modalities and are therefore not directly comparable to joint-only architectures.

4.3.3. Computational Complexity Analysis

A key advantage of DPCA-GCN is its computational efficiency. Table 3 compares model complexity, inference time, and memory usage.

Table 3. Complexity comparison. FLOPs and inference time measured on NTU RGB+D 60 with input shape $(N = 1, T = 64, V = 25, C = 3)$.

Method	Params (M)	FLOPs (G)	Time (ms)
ST-GCN [9]	3.1	16.2	8.5
2s-AGCN [10]	6.9	37.4	15.3
ST-TR [30]	5.8	45.6	22.1
CTR-GCN [14]	14.6	89.2	31.4
DPCA-GCN (Ours)	2.07	6.04	14.02

Params: number of parameters; FLOPs: floating point operations; Time: inference time per sample.

Attention Mechanism Complexity: Our chunked cross-attention achieves $\mathcal{O}(L \cdot C)$ memory complexity instead of $\mathcal{O}(L^2)$ for standard attention, where L is sequence length and C is chunk size. For typical settings ($L = 64 \times 25 = 1600$ tokens, $C = 128$), this reduces memory from $1600^2 = 2.56$M to $1600 \times 128 = 204$K elements per attention operation—a 12.5× reduction.

Multi-Modal Comparison: While multi-modal methods (CTR-GCN, InfoGCN) achieve 3-4% higher accuracy, they process four modalities (Joint, Bone, Joint Motion, Bone Motion) requiring 4× forward passes and 4× model parameters. DPCA-GCN achieves 88.72% accuracy with a single modality, offering superior efficiency for resource-constrained deployment scenarios.

4.3.4. Architectural Advantages and Analysis

The effectiveness of DPCA-GCN stems from three key architectural innovations that address fundamental limitations in existing approaches:

Specialized Dual-Path Processing: Unlike conventional approaches that process spatial and temporal information sequentially, our dual-path architecture enables specialized processing for each modality. The spatial pathway effectively captures intra-frame joint relationships through graph convolution and spatial transformers, while the temporal pathway models inter-frame dynamics using temporal convolution and temporal transformers. This separation allows each pathway to optimize for domain-specific characteristics while maintaining system coherence.

Memory-Efficient Cross-Attention: Our chunked cross-attention mechanism addresses practical memory constraints while providing regularization benefits. By achieving $O(L \cdot C)$ complexity instead of $O(L^2)$ required by standard transformer approaches, the model can process longer sequences without memory limitations. The bidirectional information exchange enables spatial features to attend to temporal features and vice versa, capturing complex spatio-temporal correlations missed by sequential processing methods.

Adaptive Feature Integration: The learnable gating mechanism automatically determines optimal balance between spatial and temporal features for different actions. This dynamic adaptation proves particularly beneficial for actions with varying spatial–temporal characteristics, as evidenced by consistent performance across diverse evaluation protocols.

The efficiency–accuracy trade-off analysis demonstrates that while multi-modal methods achieve 3–4% higher accuracy through 4× computational complexity, DPCA-GCN provides exceptional value within the joint-only paradigm. Our exceptional top-5 performance (99.14% on cross-view evaluation) indicates that the model effectively captures the most relevant action patterns, suggesting strong practical applicability for real-world deployment scenarios where computational efficiency is crucial.

4.4. Ablation Studies

To validate the effectiveness of each architectural component, we conduct comprehensive ablation studies on NTU RGB+D 60 X-Sub protocol. Table 4 presents the results.

Table 4. Ablation study results on NTU RGB+D 60 X-Sub. All variants use identical training configurations (80 epochs, batch size 16, learning rate 0.001) for fair comparison.

Variant	Configuration	Top-1 (%)	Top-5 (%)	Δ Accuracy
Full DPCA-GCN	**All components enabled**	**88.72**	**96.97**	**Baseline**
No Cross-Attention	Simple concatenation	87.95	97.32	−0.77
Unidirectional Attention	Spatial→Temporal only	87.81	97.26	−0.91
No Adaptive Gate	Fixed $\alpha = \beta = 0.5$	86.27	96.86	−2.45
Single-Path	Unified ST processing	84.59	96.18	−4.13

4.4.1. Impact of Core Components

Cross-Attention Mechanism: We compare our bidirectional cross-attention with a baseline that simply concatenates spatial and temporal features without attention-based fusion. Removing cross-attention (No Cross-Attention variant) results in 87.95% accuracy, demonstrating that explicit cross-modal information exchange is crucial. Without cross-attention, spatial and temporal pathways operate independently until final fusion, failing to capture their intricate dependencies.

Bidirectional vs. Unidirectional Attention: We evaluate unidirectional cross-attention where only spatial features attend to temporal features ($\mathbf{A}_{s \to t}$ only), mimicking asymmetric fusion strategies in prior work. The Unidirectional variant achieves 87.81% accuracy.

The 0.91% drop compared to bidirectional attention confirms that symmetric information exchange is beneficial—temporal features also benefit from attending to spatial patterns, not just vice versa.

Adaptive Gate: Disabling the adaptive gate and using fixed fusion weights ($\alpha = \beta = 0.5$) yields 86.27% accuracy (No Adaptive Gate). This 2.45% degradation validates our hypothesis that action-specific spatial–temporal balancing improves recognition, as different actions benefit from different modality emphases.

Dual-Path Architecture: We implement a Single-Path baseline that processes spatial and temporal information jointly through a single unified pathway (similar to standard ST-GCN architecture). This variant achieves 84.59%, confirming that explicit pathway separation enables modality-specific learning that improves performance.

4.4.2. Chunk Size Analysis

Our chunked attention mechanism introduces a trade-off between memory efficiency and attention granularity. Table 5 presents results across different chunk sizes.

Table 5. Impact of chunk size on accuracy and computational efficiency.

Chunk Size	Top-1 (%)	Time (ms)	Complexity
32	86.29	8.75	$\mathcal{O}(L \cdot 32)$
64	88.55	10.5	$\mathcal{O}(L \cdot 64)$
128 (default)	88.72	14.02	$\mathcal{O}(L \cdot 128)$

Observations: Our analysis of different chunk sizes reveals the following: (1) chunk sizes 64 and 128 achieve similar accuracy, indicating that moderate chunking does not sacrifice model capacity; and (2) smaller chunks (32) underperform due to their limited attention scope within chunks.

4.5. Limitations and Failure Case Analysis

While DPCA-GCN demonstrates competitive performance, several limitations warrant discussion:

Performance Gap vs. Multi-Modal Methods: Our joint-only approach achieves 88.72% on NTU 60 X-Sub, while multi-modal methods like InfoGCN reach 93.0%—a 4.3% gap. This highlights an inherent limitation: single-modality processing cannot fully capture the richness of multi-modal representations. Bone structure and motion velocity provide complementary cues that joint coordinates alone cannot encode. Failure cases : predominantly occur in action pairs with similar joint trajectories but different motion speeds (e.g., "walking" vs. "running") or fine-grained hand manipulations where bone-level features are more discriminative than joint positions.

Long-Range Temporal Dependencies: Despite chunked attention improving efficiency, our model processes sequences of 64 frames due to memory constraints. Actions spanning longer durations may have critical temporal patterns outside this window. Failure analysis : reveals that our model struggles with actions having extended preparation phases (e.g., "taking off jacket"—the critical discriminative motion occurs after initial frames that resemble "wearing jacket").

Computational Cost vs. Lightweight Models: While more efficient than multi-modal methods, DPCA-GCN is still heavier than lightweight alternatives like Shift-GCN [13] due to transformer components.

Confusion Matrix and Misclassification Analysis: We perform detailed per-class analysis to identify systematic failure modes. The key findings are visualized in Figures 6 and 7 . The most frequently misclassified action pairs and their underlying causes are as follows.

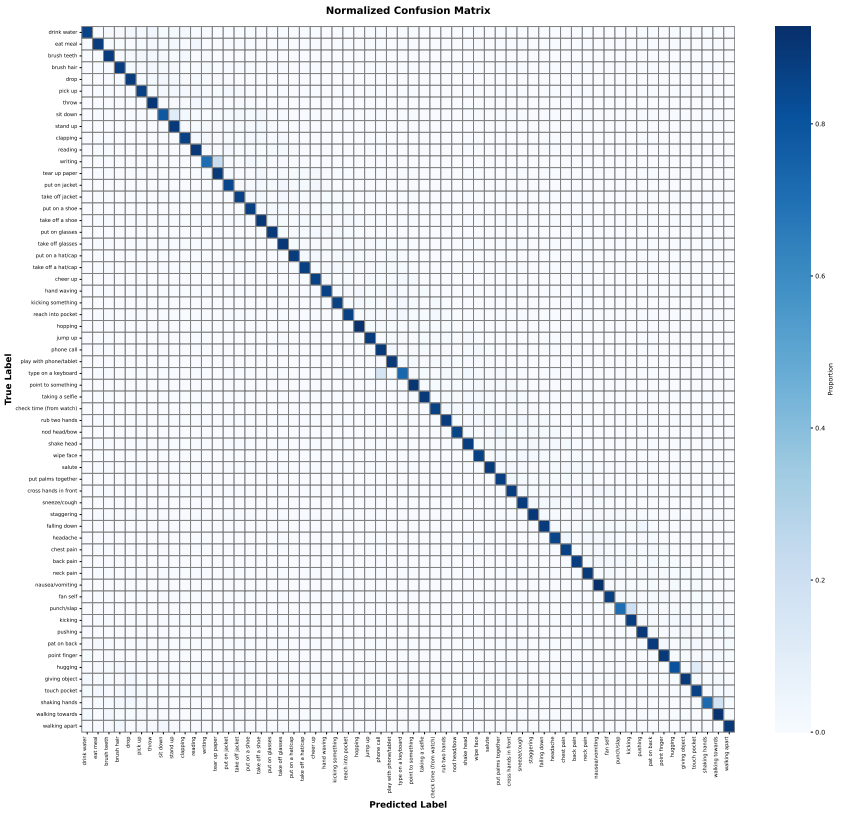


Figure 6. Normalized confusion matrix for DPCA-GCN on the NTU RGB+D 60 X-Sub dataset. Rows represent true labels and columns represent predicted labels. Diagonal elements indicate correct classifications, while off-diagonal elements show misclassifications, with darker shades indicating higher confusion rates. This matrix provides a comprehensive overview of per-class performance and inter-class confusions.

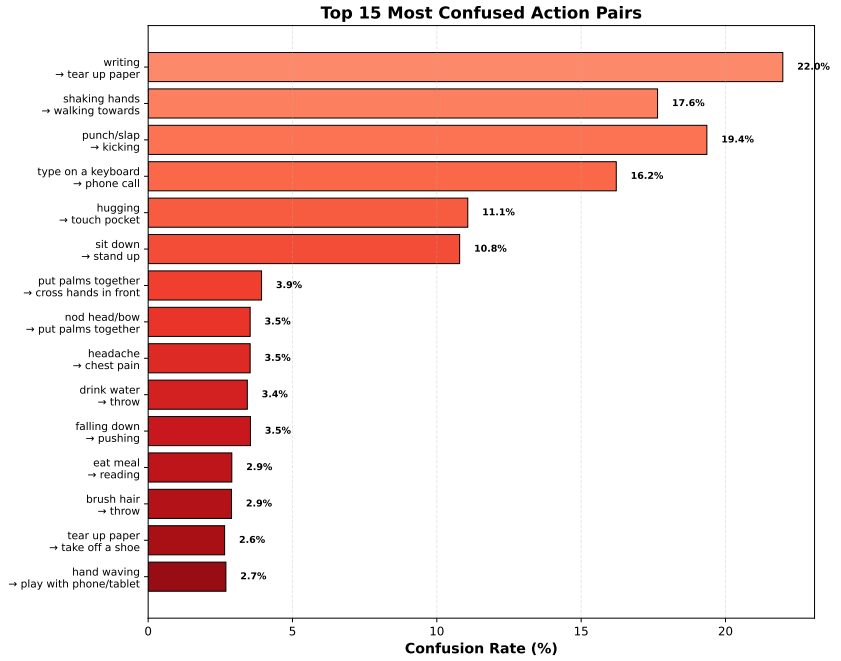


Figure 7. Top 15 most frequently misclassified action pairs by DPCA-GCN on the NTU RGB+D 60 X-Sub dataset. The confusion rates highlight systematic challenges in distinguishing similar actions, such as "writing" vs. "tear up paper" (22.0%) and "shaking hands" vs. "walking towards" (17.6%), which often share similar movement trajectories or spatial configurations.

The misclassification patterns are further analyzed across key action categories. Table 6 highlights the most frequent confusion pairs, while specific insights are drawn from detailed category analysis:

Interaction Actions (avg. 84.02% accuracy): This category, including actions like handshaking, hugging, punching, and kicking, often involves subtle inter-person dynamics or similar arm trajectories. Shaking hands vs. walking towards: 17.6% confusion rate—both involve similar arm trajectories, and distinguishing requires modeling subtle interaction patterns. *Hugging vs. touch pocket*: 11.1% confusion—spatial proximity patterns can be similar; temporal dynamics and specific contact points are key differentiators. *Punch/slap vs. kicking*: 19.4% confusion—these aggressive actions can share similar preparatory or follow-through motions.

Fine-Grained Hand Manipulations (avg. 85.61% accuracy): Actions requiring precise hand movements, such as writing or typing, present specific challenges. *Writing vs. tear up paper*: 22.0% confusion—both show similar hand-to-surface or hand–object manipulation, and without finer joint details (25-joint skeleton only includes wrist), our model struggles to distinguish grip and tool use patterns. *Type on a keyboard vs. phone call*: 16.2% confusion—repetitive hand motions can appear visually similar, especially without explicit object interaction cues.

Gross Motor Patterns (avg. 87.84% accuracy): This category encompasses actions involving larger body movements like sitting, standing, walking, and running. *Sit down vs. stand up*: 10.8% confusion—spatial endpoint states are similar; temporal direction (forward vs. backward time) is the primary distinguisher. *Walking vs. running*: (This common confusion is not explicitly in the top 20 list but is a general challenge for gross motor patterns where velocity is key). Misclassifications here often stem from subtle velocity differences that joint positions alone encode weakly.

Single-Person Daily Activities (avg. 87.81% accuracy): This broad category covers a wide range of everyday actions. Misclassifications here are often subtle and specific, requiring finer distinction of context and interaction with objects. *Drink water vs. throw*: 3.4% confusion—actions with shared initial poses but divergent subsequent movements can be challenging. *Eat meal vs. reading*: 2.9% confusion—these involve static or repetitive upper body movements that can appear similar without rich object context.

Table 6. Top 20 most frequently confused action pairs by DPCA-GCN on NTU RGB+D 60 X-Sub.

True Action	Predicted As	Count	Conf. %
writing	tear up paper	62	22.0
shaking hands	walking towards	51	17.6
punch/slap	kicking	48	19.4
type on a keyboard	phone call	42	16.2
hugging	touch pocket	32	11.1
sit down	stand up	30	10.8
put palms together	cross hands in front	11	3.9
nod head/bow	put palms together	10	3.5
headache	chest pain	10	3.5
drink water	throw	9	3.4
falling down	pushing	9	3.5
eat meal	reading	8	2.9
brush hair	throw	8	2.9
tear up paper	take off a shoe	8	2.6
hand waving	play with phone/tablet	8	2.7
shake head	staggering	8	3.0
sneeze/cough	headache	8	3.1
walking apart	brush hair	8	3.0
eat meal	stand up	7	2.5
brush teeth	pick up	7	2.6

5. Conclusions

In this work, we introduced DPCA-GCN, a novel dual-path cross-attention framework for skeleton-based human action recognition that addresses fundamental limitations in existing approaches. Our key innovations include a dual-path architecture that processes spatial and temporal information through specialized pathways, a memory-efficient cross-attention mechanism that enables effective cross-modal feature fusion, and adaptive gating for optimal feature combination.

Extensive experiments on NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that DPCA-GCN achieves competitive performance with joint-only accuracies of 88.72%/94.31% and 82.85%/83.65%, respectively, while maintaining significantly lower computational complexity compared to multi-modal approaches. The exceptional top-5 performance indicates strong practical applicability for real-world deployment scenarios.

Building upon DPCA-GCN's strengths, future work will explore enhancing its capabilities across several promising avenues. Primarily, we aim to extend the dual-path framework to seamlessly integrate additional modalities, such as bone and motion data, thereby narrowing the performance gap with state-of-the-art multi-modal methods while rigorously preserving architectural efficiency. Concurrently, we plan to investigate advanced dynamic graph construction techniques, allowing the model to adapt its understanding of joint relationships based on the specific action context, and to develop hierarchical temporal modeling pathways capable of capturing both fine-grained, short-term movements and the broader, long-range structure of complex actions. These directions promise to further unlock DPCA-GCN's potential for robust and nuanced action recognition.

Author Contributions: Conceptualization, K.L. and J.R.; Methodology, K.L.; Software, K.L. and J.R.; Validation, K.L. and J.R.; Formal analysis, K.L. and J.R.; Investigation, K.L., A.M.M. and J.R.; Resources, K.L. and K.E.F.; Data curation, K.L.; Writing—original draft, K.L.; Writing—review and editing, K.L., K.E.F., A.M.M., H.T. and J.R.; Visualization, K.L.; Supervision, J.R.; Project administration, J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets analyzed in this study are publicly available benchmark datasets (NTU RGB+D 60 and NTU RGB+D 120) as referenced in the manuscript. The datasets are available at https://rose1.ntu.edu.sg/dataset/actionRecognition/ (accessed on 1 June 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DPCA-GCN	Dual-Path Cross-Attention Graph Convolutional Network
GCN	Graph Convolutional Network
HAR	Human Action Recognition
ST-GCN	Spatial–Temporal Graph Convolutional Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
MLP	Multi-Layer Perceptron
GAP	Global Average Pooling

References

1. Wang, C.; Yan, J. A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. *IEEE Access* **2023**, *11*, 59562–59584. Available online: https://ieeexplore.ieee.org/document/10143178 (accessed on 4 December 2025). [CrossRef]
2. Zhang, H.; Zhang, Y.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.; Chen, D. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2023**, *23*, 1005. [CrossRef] [PubMed]

3. Zhao, H.; Torralba, A.; Gupta, A. Making the Invisible Visible: Action Recognition Through Walls and Occlusions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
4. Ren, B.; Liu, M.; Ding, R.; Liu, H. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. *Cyborg Bionic Syst.* **2024**, *5*, 0100. [CrossRef] [PubMed]
5. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 914–927. [CrossRef] [PubMed]
6. Xia, L.; Chen, C.-C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
7. Vishwakarma, D.K.; Kapoor, R. Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Syst. Appl.* **2015**, *42*, 6957–6965. [CrossRef]
8. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
9. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA , USA, 2–7 February 2018; Volume 32.
10. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
11. Ye, Y.; Liu, P.; Wang, X. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 55–63.
12. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
13. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 183–192.
14. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13339–13348.
15. Chi, N.; Hou, Z.; Xu, H.; Song, J. Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), New Orleans, LA, USA, 18–24 June 2022; pp. 20154–20164.
16. Kilic, U.; Oztimur Karadag, O.; Tumuklu Ozyer, G. AGMS-GCN: Attention-Guided Multi-Scale Graph Convolutional Networks for Skeleton-Based Action Recognition. *Knowl.-Based Syst.* **2025**, 113045. [CrossRef]
17. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv* **2020**, arXiv:2006.04768. [CrossRef]
18. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
19. Feng, D.; Wu, Z.; Zhang, J.; Ren, T. Multi-scale temporal graph neural network for skeleton-based action recognition. *Pattern Recognit.* **2023**, *135*, 109173. [CrossRef]
20. Liu, D.; Chen, P.; Yao, M.; Lu, Y.; Cai, Z.; Tian, Y. TSGCNeXt: Dynamic-static graph convolution for efficient skeleton-based action recognition. *arXiv* **2023**, arXiv:2301.10111. [CrossRef]
21. Zhou, Y.; Yan, X.; Cheng, Z.-Q.; Yan, Y.; Dai, Q.; Hua, X.-S. BlockGCN: Redefine topology awareness for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , Seattle, WA, USA, 16–22 June 2024; pp. 2049–2058.
22. Myung, W.; Su, N.; Xue, J.; Wang, G. Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Trans. Image Process.* **2024**, *33*, 2477–2490. [CrossRef]
23. Li, Y.; Li, N. DF-GCN: A Lightweight Dynamic Dual-Stream Graph Convolutional Network for Skeleton-Based Action Recognition. In Proceedings of the IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA), Shenyang, China, 28–30 June 2025; pp. 2181–2186.
24. Chen, D.; Chen, M.; Wu, P.; Wu, M.; Zhang, T.; Li, C. Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. *Sci. Rep.* **2025**, *15*, 4982. [CrossRef] [PubMed]
25. Lasri, K.; Riffi, J.; Fazazy, K.E.; Mahraz, A.M.; Tairi, H. Hybrid attention-inflated 3D architecture for human action recognition. *Multimed. Tools Appl.* **2025**, *84*, 42757–42776. [CrossRef]
26. Ramanathan, M.; Yau, W.-Y.; Magnenat-Thalmann, N.; Teoh, E.K. Mutually reinforcing motion-pose framework for pose invariant action recognition. *Int. J. Biom.* **2019**, *11*, 113–147. [CrossRef]

27. Lasri, K.; Riffi, J.; Fazazy, K.E.; Mahraz, A.M.; Tairi, H. Comparative Study of I3D and SlowFast Networks for Spatial-Temporal Video Action Recognition. In Proceedings of the 4th International Conference on Advances in Communication Technology and Computer Engineering (ICACTCE'24), Meknes, Morocco, 29–30 November 2024; Springer: Cham, Switzerland, 2025; pp. 435–447.
28. Plizzari, C.; Cannici, M.; Matteucci, M. Spatial Temporal Transformer Network for Skeleton-Based Action Recognition. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges (ICPR 2021), Virtual, 10–15 January 2021; Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 12663, pp. 651–664. [CrossRef]
29. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
30. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* **2021**, *208*, 103219. [CrossRef]
31. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action Recognition. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December 2020; pp. 38–54. [CrossRef]
32. Ahn, D.; Kim, S.; Hong, H.; Ko, B.C. STAR-Transformer: A Spatio-Temporal Cross-Attention Transformer for Human Action Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 3319–3328. [CrossRef]
33. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 813–824.
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
35. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 6804–6815. [CrossRef]
36. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
37. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
38. Duan, H.; Wang, J.; Chen, K.; Lin, D. DG-STGCN: Dynamic spatial-temporal graph convolutional network for skeleton-based action recognition. *arXiv* **2022**, arXiv:2210.05895. [CrossRef]
39. Wang, J.; Bergeret, E.; Falih, I. Skeleton-Based Action Recognition with Spatial-Structural Graph Convolution. *arXiv* **2024**. [CrossRef]
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
41. Chen, T.; Xu, B.; Zhang, C.; Guestrin, C. Training deep nets with sublinear memory cost. *arXiv* **2016**, arXiv:1604.06174. [CrossRef]
42. Shahroudy, A.; Liu, J.; Ng, T.-T.; Wang, G. 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), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
43. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.-Y.; Kot, A.C. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2684–2701. [CrossRef]
44. Xing, Y.; Zhu, J. Deep Learning-Based Action Recognition with 3D Skeleton: A Survey. *CAAI Trans. Intell. Technol.* **2021**. [CrossRef]
45. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal LSTM with trust gates for 3D human action recognition. In Proceedings of the European Conference Computer Vision—ECCV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 816–833.
46. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1963–1978. [CrossRef]
47. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5323–5532.
48. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2636–2645. Available online: https://openaccess.thecvf.com/content_CVPR_2020/html/Yu_BDD100K_A_Diverse_Driving_Dataset_for_Heterogeneous_Multitask_Learning_CVPR_2020_paper.html (accessed on 4 December 2025).

49. Song, Y.-F.; Zhang, Z.; Wang, L. Richly activated graph convolutional network for action recognition with incomplete skeletons. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1–5.
50. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1112–1121.
51. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2117–2126.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.