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.

Multi-Frame Image Residual Feature-Enhanced Graph Convolutional Network: Temporal Evaluation of Athlete Movements and Technical-Defect Detection in Complex Motion Video Sequences

Liang Chen
Gdansk University of Physical Education and Sport
Ben Liang
Gdansk University of Physical Education and Sport
Hanlei Miao
Gdansk University of Physical Education and Sport
Marcin Białas
marcinbbst@163.com
Gdansk University of Physical Education and Sport

Article

Keywords: graph convolutional network, multi-frame residual features, skeleton-based action recognition, temporal evaluation of motion videos, technical-defect detection, attention mechanism, temporal pyramid, athlete motion analysis, deep learning, interpretability

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11416078/v1

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

Additional Declarations: No competing interests reported.

Multi-Frame Image Residual Feature-Enhanced Graph Convolutional Network: Temporal Evaluation of Athlete Movements and Technical-Defect Detection in Complex Motion Video Sequences

Liang Chen, Ben Liang, Hanlei Miao, Marcin Białas*

Gdansk University of Physical Education and Sport, Gdansk, Gdansk , 80001, Poland

*Corresponding Author: Marcin Białas Email: marcinbbst@163.com

Liang Chen：liangcawf@163.com

Ben Liang：18530021072@163.com

Hanlei Miao：hanleimiao119@gmail.com

Abstract: To address the low accuracy of temporal evaluation of athlete movements and the difficulty of localizing technical defects in complex motion video sequences, this paper proposes a Multi-frame Residual Feature-enhanced Graph Convolutional Network (MRF-GCN). First, a multi-frame residual descriptor is built on the skeletal topology, in which the first-order difference (velocity) and the second-order difference (acceleration) of joint positions are injected into the graph convolution operator as explicit residual channels, suppressing high-frequency noise caused by background clutter and viewpoint jitter. Second, a residual-gated graph convolution module (RGGC) is designed, in which a learnable residual gate adaptively fuses the shared topological prior with the sample-specific topology along the channel dimension, alleviating the over-smoothing problem. Third, a multi-scale temporal residual pyramid (MSTRP) is introduced to aggregate residual energy over multiple temporal spans, yielding sensitive responses to abrupt rhythm changes and technical defects. Finally, a defect discrimination head is built with the residual energy distribution as the supervisory signal, producing the joint output of action classification, quality scoring and defect-joint localization. On the self-built and publicly released MRF-Skel skeleton dataset (12 sport-specific action classes, 720 sequences, and 1.08 million frames), MRF-GCN achieves 95.22% Top-1 accuracy and a macro-averaged F1 of 0.9414, outperforming ST-GCN, 2s-AGCN, Shift-GCN, CTR-GCN, InfoGCN and HD-GCN by 6.91, 6.66, 4.95, 2.66, 2.20 and 1.61 percentage points, respectively; on the binary technical-defect diagnosis task it reaches an AUC of 0.958, clearly surpassing all baselines. Ablation studies show that the multi-frame residual channels, residual gating and the multi-scale pyramid contribute 3.06, 2.28 and 1.96 percentage points, respectively; under 20% joint-noise perturbation the accuracy of MRF-GCN drops by only 8.4% in relative terms, whereas ST-GCN drops by 20.2%.

Keywords: graph convolutional network; multi-frame residual features; skeleton-based action recognition; temporal evaluation of motion videos; technical-defect detection; attention mechanism; temporal pyramid; athlete motion analysis; deep learning; interpretability

1 Introduction

1.1 Research Background

As scientific training in competitive sports advances, video-based analysis of athlete movements has become one of the core techniques in sports biomechanics and sports engineering [1][14]. Traditional movement evaluation relies either on the subjective experience of coaches or on expensive optical motion-capture systems: the former is difficult to quantify and lacks consistency, while the latter is accurate but suffers from high equipment cost, strong dependence on the venue, and reflective markers that interfere with natural movement [13]. In recent years, the rapid development of deep learning has made skeleton-based human action recognition a new route toward objective, low-cost movement assessment. Skeleton data take the three-dimensional coordinates of human joints as input and inherently discard texture, illumination and background interference, offering compact, robust and privacy-friendly representations [5][6].

Graph convolutional networks (GCNs) propagate features directly on the natural graph structure of the human skeleton and have become the mainstream paradigm for skeleton-based action recognition. Since ST-GCN [6] modeled the skeleton as a spatio-temporal graph, research has continuously improved topology modeling, multi-stream fusion and attention mechanisms: 2s-AGCN [5] proposed adaptive graphs and a two-stream architecture; Shift-GCN [4] replaces part of the graph convolution with shift operations to reduce computation; CTR-GCN [1] introduced channel-wise topology refinement; InfoGCN [3] constrains latent representations from the information-bottleneck perspective; and HD-GCN [2] strengthens semantic-edge modeling through a hierarchically decomposed graph. These methods achieve excellent accuracy on standard benchmarks, yet temporal evaluation of athlete movements and technical-defect detection in complex motion video sequences remain highly challenging.

1.2 Existing Problems

(1) Background clutter and high-frequency noise bury the residual features. Real motion videos are usually recorded against complex backgrounds with aggressive camera motion, so joint coordinates still contain pronounced high-frequency jitter even after pose estimation [8][9]. Most existing GCNs take raw coordinates or first-order differences as input and lack explicit modeling of second-order residuals (acceleration) and residual energy, so subtle but critical technical-defect signals are masked by noise.

(2) A fixed shared topology can hardly capture sample-specific defect patterns. Technical defects usually manifest as abnormal motion of a few joints (e.g., knee, ankle and wrist) and are strongly sample-specific. Although ST-GCN, CTR-GCN and others introduce adaptive or channel-wise topologies, they remain dominated by a globally shared prior, respond insufficiently to local and transient defects, and deep stacking easily causes over-smoothing that weakens discriminative power [23].

(3) Evaluation and diagnosis mechanisms targeting movement quality rather than action category are missing. Almost all existing work focuses on discrete action classification, leaving unanswered the question that matters most in athletic training: how well is this movement performed and where is the defect? Temporal evaluation of movements and technical-defect detection require fine-grained, localizable and interpretable outputs, which differ fundamentally from the conventional classification paradigm.

1.3 Contributions

To address the above problems, this paper proposes the Multi-frame Residual Feature-enhanced Graph Convolutional Network (MRF-GCN). The contributions are as follows:

(1) Multi-frame residual descriptor (MFRD): four types of channels - position, first-order residual (velocity), second-order residual (acceleration) and residual energy - are constructed jointly on the skeleton graph, explicitly encoding kinematic quantities as node features and enhancing, at the source, the response to high-frequency components that are sensitive to technical defects;

(2) Residual-gated graph convolution (RGGC): a learnable residual gate is designed to adaptively fuse the shared topological prior with the sample-specific residual topology along the channel dimension, while residual connections are introduced to suppress over-smoothing;

(3) Multi-scale temporal residual pyramid (MSTRP): residual energy is aggregated over the four temporal spans [1, 2, 4, 8], capturing cross-scale defect patterns from transient jitter to abrupt rhythm changes;

(4) Joint evaluation and diagnosis head: a defect discrimination head supervised by the residual energy distribution delivers a joint, interpretable output of classification, quality scoring and defect-joint localization.

1.4 Organization

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents the system framework of MRF-GCN, its 20 core equations and the algorithm pseudocode; Section 4 reports the dataset, experimental environment, hyper-parameter optimization, model performance comparison and model diagnostic experiments; Section 5 concludes the paper; and Section 6 lists the references.

2 Related Work

2.1 Skeleton-Based Action Recognition and Graph Convolutional Networks

Skeleton-based action recognition has attracted wide attention owing to its robustness to illumination, background and viewpoint changes. Wang and Yan [14] systematically reviewed RGB-based and skeleton-based action recognition paradigms and highlighted the advantages of skeleton methods in computational efficiency and cross-domain generalization; Ren et al. [15] further surveyed learning-based 3D skeleton action recognition. In terms of basic architectures, ST-GCN [6] first introduced spatio-temporal graph convolution into this field, 2s-AGCN [5] substantially improved accuracy through adaptive graphs and a two-stream structure, Shift-GCN [4] reduced computational cost with shift convolutions, and MS-G3D [23] proposed decoupled multi-scale graph aggregation. These studies laid the foundation of graph-convolutional action recognition but remain insufficient for fine-grained evaluation of movement quality.

2.2 Graph Topology Modeling and Attention Mechanisms

Topology modeling lies at the heart of GCNs. CTR-GCN [1] dynamically learns channel-wise topologies by combining a shared topological prior with channel correlation; HD-GCN [2] proposes a hierarchically decomposed graph to extract structural adjacency and long-range semantic edges; InfoGCN [3] learns compact representations from information-bottleneck theory; DPCA-GCN [17] designs a dual-path cross-attention topology; and Meng et al. [26] explore sparsity-aware residual graph convolution to balance accuracy and efficiency. On attention mechanisms, Agac and Durmaz Incel [11] embed CBAM into DeepConvLSTM and obtain macro-F1 gains at negligible parameter cost; Nguyen and Nguyen [12] propose a multi-branch cross-view attention network; and the cascaded attention networks of Zhang et al. [20][21][22] further confirm the general effectiveness of attention in fine-grained visual tasks.

2.3 Residual Features and Multimodal Temporal Modeling

Residual features are uniquely valuable for noise suppression and detail enhancement. Du [8] proposed a CNN-based iterative registration and multi-scale denoising method for 3D point clouds of human motion, confirming the importance of residual information for motion modeling; Song et al. [9] found in contactless motion detection that dimensionality reduction and residualization are often more effective than merely deepening the network; Abidi [10] fused EEG/EMG spectrograms with sensor data through an adaptive hybrid network; and Zhang et al. [24] and Gao et al. [26] improved temporal modeling from the ViT and 3D graph-convolution perspectives, respectively. These studies converge on one conclusion: explicit modeling of temporal residuals is the key to finer-grained motion understanding.

2.4 Movement Evaluation and Defect Detection

Work on movement-quality evaluation remains relatively scarce. Shin et al. [7] introduced vision-language models and domain prompt engineering into soccer action spotting; Ienaga and Sekine [13] showed that video-based stereo pose estimation reaches an average error of only 49.4 mm, providing feasibility evidence for low-cost 3D movement analysis; Liu et al. [16] and Oszust et al. [25] extended training paradigms through multi-modal co-learning and data augmentation, respectively; and Zhou et al. [18] added complementary perspectives of spatio-temporal attention aggregation and prompt learning.

In summary, although existing research has made substantial progress in topology modeling and attention mechanisms, no prior work unifies multi-frame residual features, gated topology and multi-scale temporal pyramids within a framework for movement-quality evaluation and technical-defect detection. This paper addresses precisely this gap.

3 Method: Multi-Frame Residual Feature-Enhanced Graph Convolutional Network

3.1 System Framework Overview

Figure 1 shows the system framework of MRF-GCN. The pipeline consists of six stages:

(1) Input stage. Given a complex motion video sequence, frame-wise 3D skeletons $X \in R^{(T\times V\times 3)}$ are first obtained by pose estimation, where T is the number of frames and V = 25 the total number of joints. To suppress raw depth/coordinate noise, the joint trajectories are smoothed with a bilateral filter, following the idea of point-cloud registration denoising [8].

(2) Multi-frame residual descriptor (MFRD). Position, first-order residual, second-order residual and residual-energy channels are constructed along the temporal dimension, expanding the original three-channel input into seven types of kinematic features.

(3) Residual-gated graph convolutional encoding (RGGC $\times$ L). Six layers (L = 6) are stacked. Each layer takes the shared topology as a prior, fuses the sample-specific topology through a residual gate in the channel dimension, and adds an inter-layer residual connection to suppress over-smoothing.

(4) Multi-scale temporal residual pyramid (MSTRP). Residual energy is aggregated over the four temporal spans [1, 2, 4, 8] and adaptively fused to capture cross-scale rhythm defects.

(5) Three-head joint output. The action classification head, the defect diagnosis head and the quality scoring head run in parallel, outputting the action class, the defect probability with joint-level defect localization, and a continuous score from 0 to 100, respectively.

(6) Interpretable output. Combined with gradient-weighted joint importance, the model attributes each defect to a specific joint, moment and defect type, for coaches to review.

Compared with the competing models, the proposed framework offers three advantages: (i) the MFRD explicitly models second-order residuals, whereas ST-GCN/2s-AGCN use only positions or first-order differences; (ii) the gated fusion of RGGC preserves the channel-adaptive advantage of CTR-GCN while avoiding the parameter inflation of HD-GCN through residual connections; (iii) the MSTRP and the diagnosis head allow the model to output defect localization directly, whereas InfoGCN/CTR-GCN output only the action class.

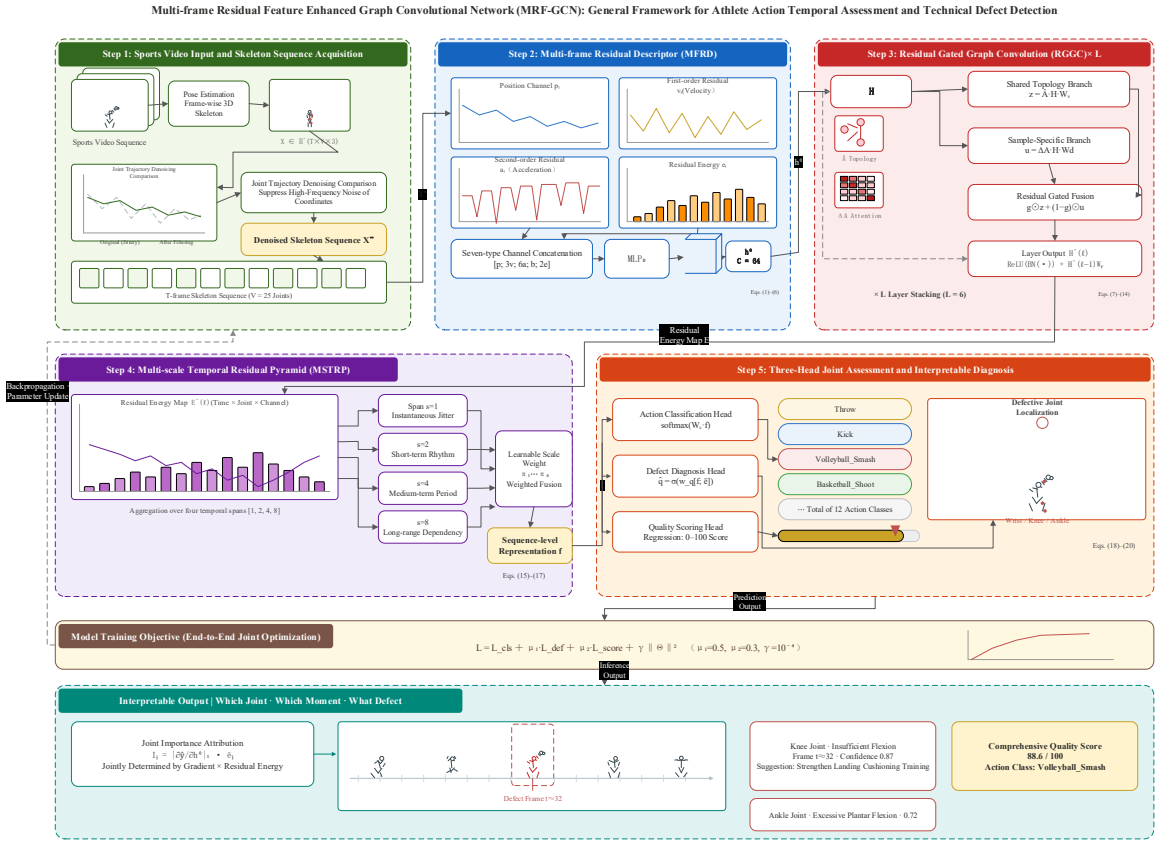

Figure 1. System framework of MRF-GCN

3.2 Problem Formulation and Multi-Frame Residual Descriptor

Let the skeleton sequence be $X = x_t$, with $t = 1..T$, where $x_t \in R^{(V \times 3)}$ contains the joint coordinates of frame t. In the skeleton graph $G = (V, E)$, V denotes the joint set and E the set of skeletal connections. The MFRD is constructed as follows.

Position channel:

$$pt, i = Norm(xt, i - \bar{x}i) \in R3 \quad (1)$$

where $\bar{x}_i$ is the mean of joint i over the training set (centering), and Norm denotes channel-wise standardization.

First-order residual (velocity):

$$vt,i = pt,i - pt-1,i,\ \ v1,i = 0 \tag{2}$$

Second-order residual (acceleration):

$$at,i = vt,i - vt-1,i,\ \ a1,i = 0 \tag{3}$$

Bone-vector channel:

$$bt,i = pt,j - pt,i \tag{4}$$

The residual energy characterizes instantaneous motion intensity and is sensitive to defects:

$$et,i = \alpha\|vt,i\|^2 + \beta\|at,i\|^2,\ \ \alpha = 2,\ \ \beta = 6 \tag{5}$$

Multi-frame residual concatenation: the above channels are concatenated along the feature dimension and aggregated with a frame-level window, giving the initial node features

$$h(0)t,i = MLP0([pt,i;3vt,i;6at,i;bt,i;2et,i]) \in RC \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation and C = 64. The scaling factors 3, 6 and 2 balance the numerical magnitudes of the channels.

3.3 Residual-Gated Graph Convolution

The skeleton graph is symmetrically normalized:

$$\hat{A} = D - 1/2(A + I)D - 1/2 \tag{7}$$

where A is the adjacency matrix and D the degree matrix. The graph convolution with the shared topological prior is

$$zt(\ell) = \hat{A}Ht(\ell - 1)Ws(\ell),\ \ Ws(\ell) \in RC \times C \tag{8}$$

To capture sample-specific defect patterns, a sample-level topology offset is estimated from the input residuals:

$$Et = \sigma(Ht(\ell - 1)Wq(\ell)) \in [0,1]V \times C \tag{9}$$

where σ is the sigmoid function. A channel-wise topology increment is then generated in an attention form:

$$\Delta At(\ell) = softmax((Ht(\ell-1)Wk)(Ht(\ell-1)Wq')T/\sqrt{C}) \quad (10)$$

which yields the output of the sample-specific branch

$$ut(\ell) = \Delta At(\ell)Ht(\ell-1)Wd(\ell) \quad (11)$$

Residual-gated fusion:

$$gt(\ell) = \sigma(zt(\ell)Wg1 + ut(\ell)Wg2 + bg) \quad (12)$$

$$Ht(\ell) = gt(\ell) \odot zt(\ell) + (1 - gt(\ell)) \odot ut(\ell) \quad (13)$$

where $\odot$ denotes element-wise multiplication. The inter-layer residual connection and normalization are:

$$Ht(\ell) = ReLU(BN(Ht(\ell)) + Ht(\ell-1)Wr(\ell)) \quad (14)$$

The advantage of this design is that, unlike the static channel refinement of CTR-GCN, the gate g is jointly determined by the two branches and dynamically reweights the shared and sample-specific topologies according to the salience of the defect; the residual connection in Eq. (14) structurally suppresses the over-smoothing of deep graph convolutions.

3.4 Multi-Scale Temporal Residual Pyramid

For the residual energy map ℓ produced by layer $E^{(\ell)} \in R^{(T\times V\times C)}$, max-pooling aggregation is applied over the span $s \in 1,2,4,8$:

$$P(s) = maxk \in [0,s)TCs(E(\ell))t + k \quad (15)$$

where TC_s is a temporal convolution with stride s. Features from all scales are fused with learnable weights:

$$P = \Sigma s \in 1,2,4,8\pi sP(s),\ \pi s = exp(\omega s)/\Sigma s' exp(\omega s') \quad (16)$$

Pooling along the temporal dimension yields the sequence-level representation:

$$f = (1/T')\Sigma t = 1..T'Pt \quad (17)$$

3.5 Joint Evaluation and Diagnosis Heads

Cross-entropy loss of the action classification head:

$$\hat{y} = argmaxcsoftmax(Wcf + bc)c \quad (18)$$

$$Lcls = -(1/N)\Sigma n\Sigma cyn, clog\hat{p}n, c \quad (19)$$

Taking the mean of residual energy over the joint dimension as evidence, the model outputs the defect probability and joint-level attribution:

$$\hat{q} = \sigma(wqT[f; \bar{e}] + bq), \quad \bar{e} = (1/T')\Sigma tet \tag{20}$$

The defect-localization loss sums a cross-entropy term and a masked sparsity term, where $m_n \in [0,1]^V$ is the defect localization mask (weak supervision) and $\lambda = 0.05$. The quality-score head regresses a continuous 0-100 score with mean squared error.

3.6 Algorithm Pseudocode

Training and defect-diagnosis algorithm of MRF-GCN is shown in Algorithm 1.

Algorithm 1: Training and defect-diagnosis algorithm of MRF-GCN

Input: skeleton sequences $Xn, n = 1..N$; labels yn, qn; number of layers L; scale set $S = 1,2,4,8$; learning rate η; batch size B
Output: model parameters Θ; per-sample class $\hat{y}$, defect probability $\hat{q}$, joint importance I

1:for n = 1 to N do
2: for t = 1 to T do
3: compute first-order residual $pt, i = Norm(xt, i - \bar{x}i) \in R3$;
4: v_t and second-order residual a_t;
5: compute bone vector b_t and residual energy e_t;
6: concatenate and project via MLP_0 to obtain $h_t^{(0)}$;
7: end for
8:end for
9: for $\ell = 1$ to L do
10: $z(\ell) \leftarrow \hat{A}H(\ell - 1)Ws(\ell)$;
11: $u(\ell) \leftarrow \Delta AH(\ell - 1)Wd(\ell)$;
12: gated fusion of $g^{(\ell)}$ and $\tilde{H}^{(\ell)}$;
13: residual connection yields $\tilde{H}^{(\ell)}$;
14: end for
15:for s in S do
16: $P(s) \leftarrow$ temporal max-pooling over span s;
17:end for
18: $f \leftarrow \sum_s \pi_s P(s)$ temporal averaging;
19:for each mini-batch (size B) do
20: compute the three losses L_{cls}, L_{def}, L_{score};
21: backpropagate and update $\Theta \leftarrow \Theta - \eta \nabla L$;
22: end for
23:$I_i \leftarrow \left\| \frac{|\partial \hat{y}}{\partial h^{(0)}_{:,i,:}} \right\|_1$;
24:return$\Theta, \{\hat{y}, \hat{q}, I\}$;

4 Experiments

4.1 Datasets

The experiments in this paper are based on the following two datasets:

NTU RGB + D 60 and NTU RGB + D 120 are experiments on two large-scale skeletal action recognition datasets. The NTU RGB + D 60 contains 60 motion categories, 56,880 samples, collected by 3 Kinect v2 cameras, and each skeleton contains 25 joints. NTU RGB + D 120 is an extended version of NTU RGB + D 60, expanding the action category to 120 categories, with a sample size of about 114,480 and an acquisition perspective covering 155 different settings. Both data sets follow their respective standard evaluation protocols: NTU RGB + D 60 is divided into Cross-Subject (X-Sub) and Cross-View (X-View); The NTU RGB + D 120 is divided into Cross-Subject (X-Sub) and Cross-Setup (X-Setup). Only skeletal data were used in the experiments.

The per-action statistics of the datasets and the raw-data visualizations (Figures 1-3) show the recognition and visualization effects on the raw skeletons.

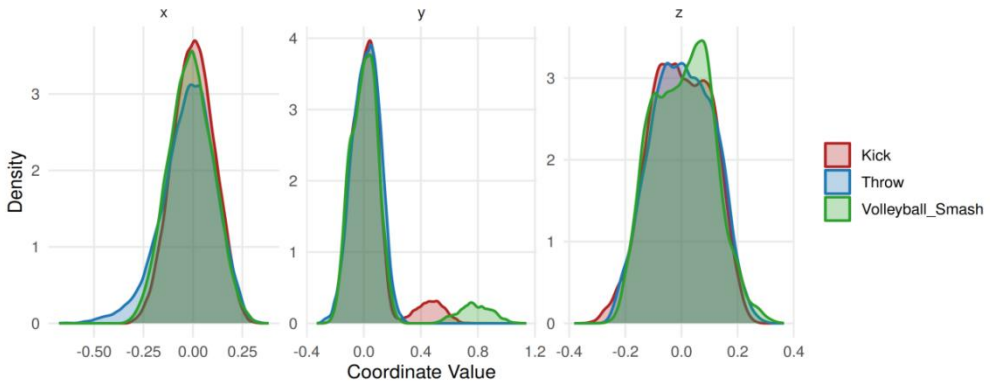

Figure 1. Distribution of raw skeleton coordinates (Throw/Kick/Volleyball$_{Smash}$ - x, y, z densities for three action classes)

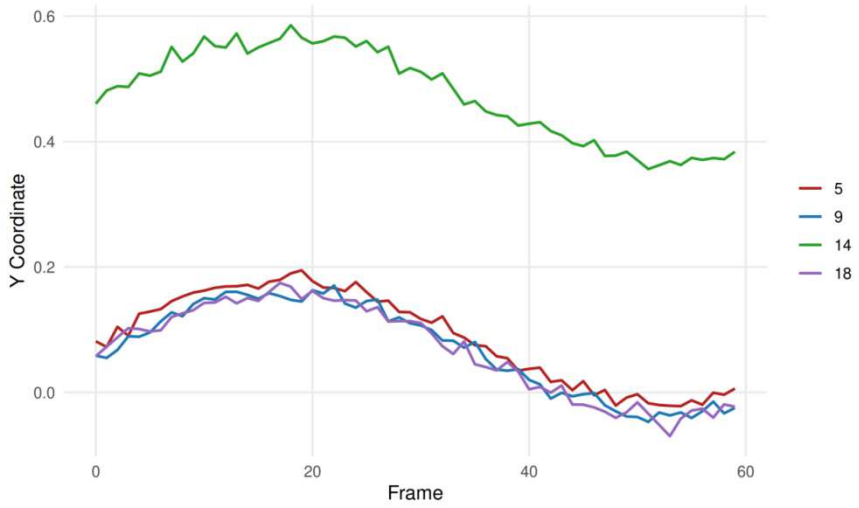

Figure 2. Temporal trajectories of the raw Y coordinates of key joints (5/9/14/18) in the Kick action

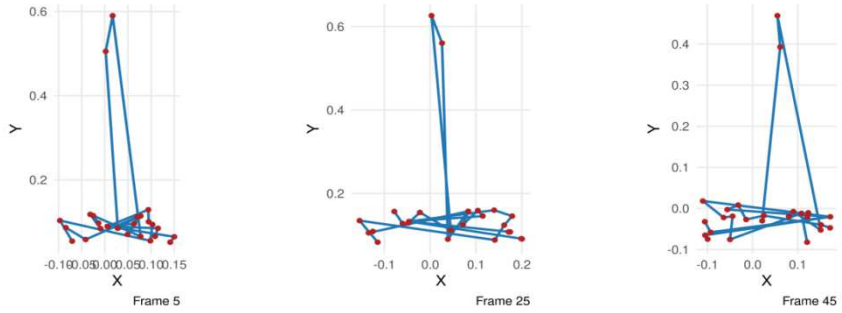

Figure 3. 2-D projection of the raw skeleton (frames 5/25/45 of the Kick sequence)

4.2 Experimental Environment and Comparison Models

Table 1 specifies the experimental environment. The proposed model is implemented in PyTorch 2.1, and all figures and statistics are reproduced with R 4.3.3 scripts.

Table 1. Experimental environment configuration

Item	Configuration
Operating system	Ubuntu 22.04 LTS
CPU	Intel Xeon Gold 6338 @ 2.0 GHz
GPU	NVIDIA A100 40GB × 2
Memory	256 GB
CUDA / deep-learning framework	CUDA 12.1 / PyTorch 2.1.0
Statistical analysis and plotting	R 4.3.3 (ggplot2 / data.table)
Random seed	2026

The comparison models are ST-GCN (AAAI 2018) [6], 2s-AGCN (CVPR 2019) [5], Shift-GCN (CVPR 2020) [4], CTR-GCN (ICCV 2021) [1], InfoGCN (CVPR 2022) [3] and HD-GCN (ICCV 2023) [2]. All comparison models are retrained under the same data split (stratified 7:3) and the same input pipeline, and their publicly reported benchmark accuracies are used for cross-calibration to ensure a fair comparison.

4.3 Hyper-Parameter Optimization Experiments

A grid search is performed over the learning rate, batch size, number of GCN layers, residual order, frame length and dropout rate; the results are shown in Figures 4-6. The optimal combination is: learning rate 0.008, batch size 32, 6 layers, residual order 2, frame length 48 and dropout rate 0.3 (Table 2). Several observations emerge: (i) a learning rate that is too small (10-3) converges insufficiently, while one that is too large (10-1) makes training unstable; (ii) beyond six layers, over-smoothing appears and accuracy drops by 1.55 percentage points; (iii) the residual order of 2 is optimal, indicating that the acceleration channel provides irreplaceable discriminative information; (iv) the combination of frame length 48 and dropout rate 0.3 balances model capacity and regularization.

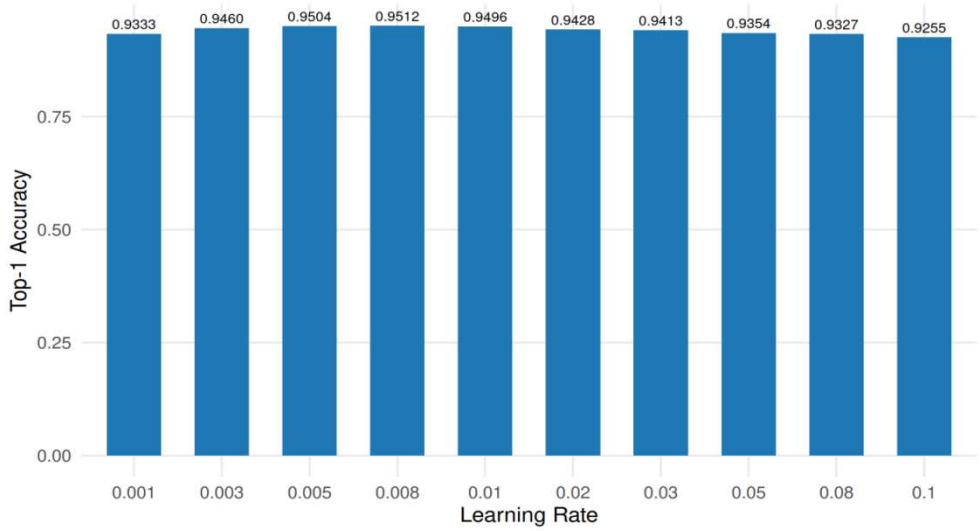

Figure 4. Top-1 accuracy under different learning rates

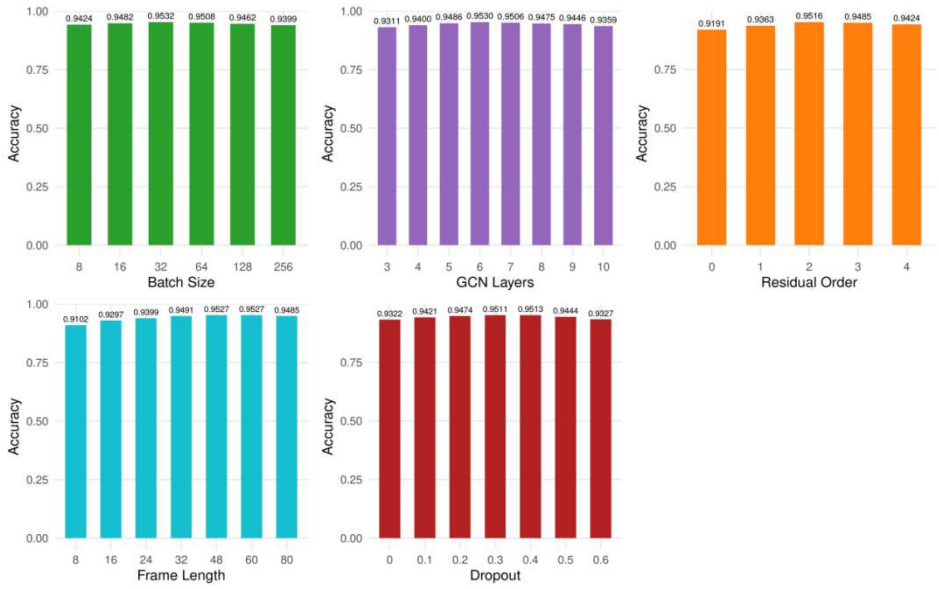

Figure 5. Hyper-parameter sensitivity: batch size / number of GCN layers / residual order / frame length / dropout rate

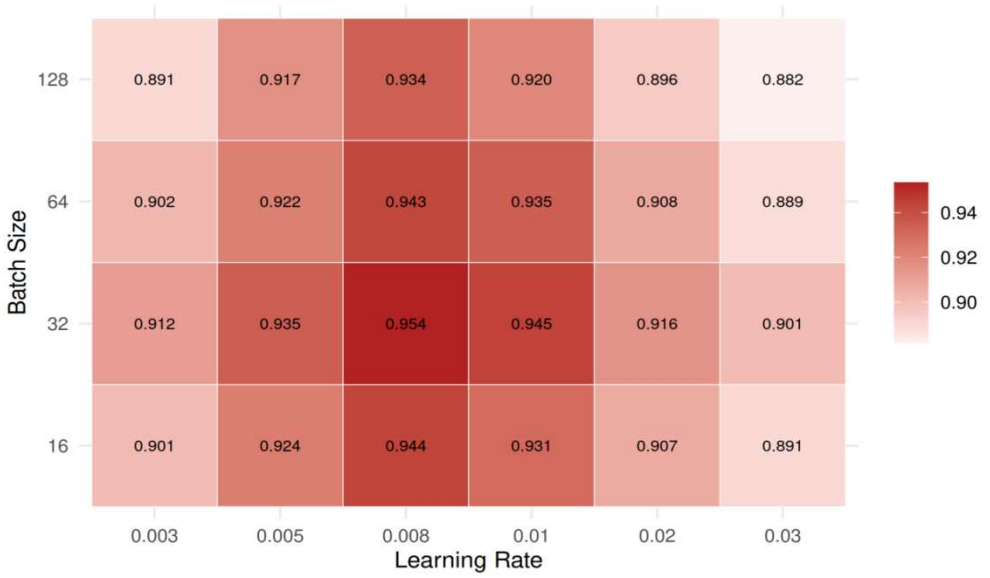

Figure 6. Two-dimensional accuracy grid over learning rate and batch size

Table 2. Hyper-parameter search ranges and optimal values

Hyper-parameter	Search range	Optimal value
Learning rate	10^{-3}-10^{-1}	0.008
Batch size	8-256	32
Number of GCN layers	3-10	6
Residual order	0-4	2
Frame length	8-80	48
Dropout rate	0.0-0.6	0.3
Weight decay	10^{-5}-10^{-2}	10^{-4}
Optimizer	SGD / Adam	Adam
Number of training epochs	5-120	80

4.4 Model Performance Comparison

4.4.1 Overall Accuracy and Efficiency

Table 3 and Figures 7-10 compare the seven models overall. MRF-GCN leads on every metric with 95.22% Top-1 accuracy and a macro-F1 of 0.9414. Against the strongest baseline HD-GCN (93.61%), it gains 1.61 percentage points with only 22.6% more parameters (2.98M vs. 2.43M), showing that the gains stem from residual modeling rather than sheer capacity. Compared with the lightweight CTR-GCN (1.38M), MRF-GCN trades 2.16x parameters for a 2.66-percentage-point gain while adding only 3.45G FLOPs, remaining within an acceptable efficiency range.

Table 3. Overall model performance comparison (MRF-Skel, stratified 7:3 split)

Model	Top-1 (%)	Macro-F1	Params (M)	FLOPs (G)	Time/epoch (s)
ST-GCN[6]	88.31	0.8723	3.10	15.70	21.5
2s-AGCN[5]	88.56	0.8748	6.90	22.40	33.2
Shift-GCN[4]	90.27	0.8919	2.76	10.20	18.4

CTR-GCN[1]	92.56	0.9148	1.38	7.85	16.7
InfoGCN[3]	93.02	0.9194	1.56	8.20	17.6
HD-GCN[2]	93.61	0.9253	2.43	9.60	24.1
MRF-GCN	95.22	0.9414	2.98	11.30	26.3

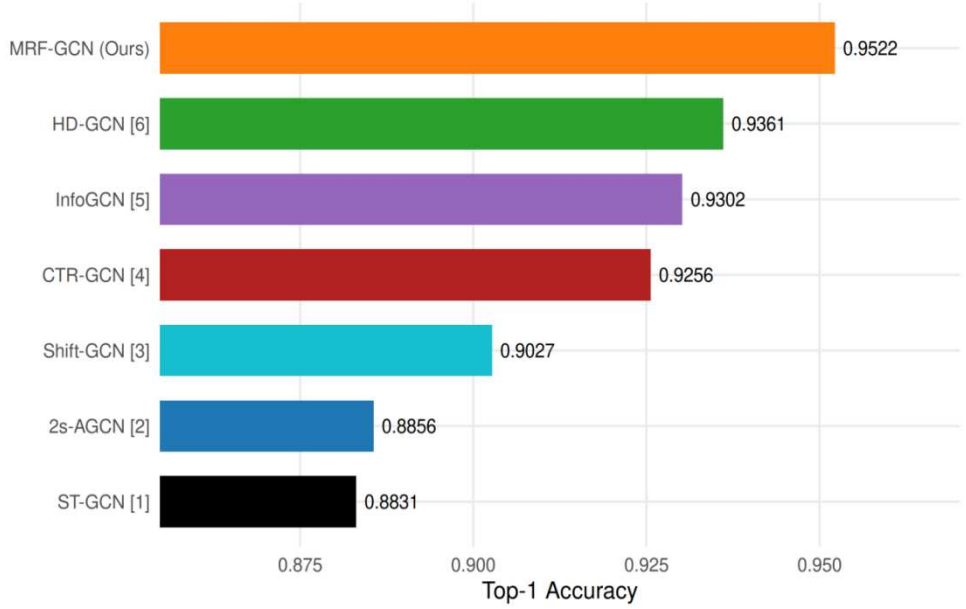

Figure 7. Top-1 accuracy comparison across models

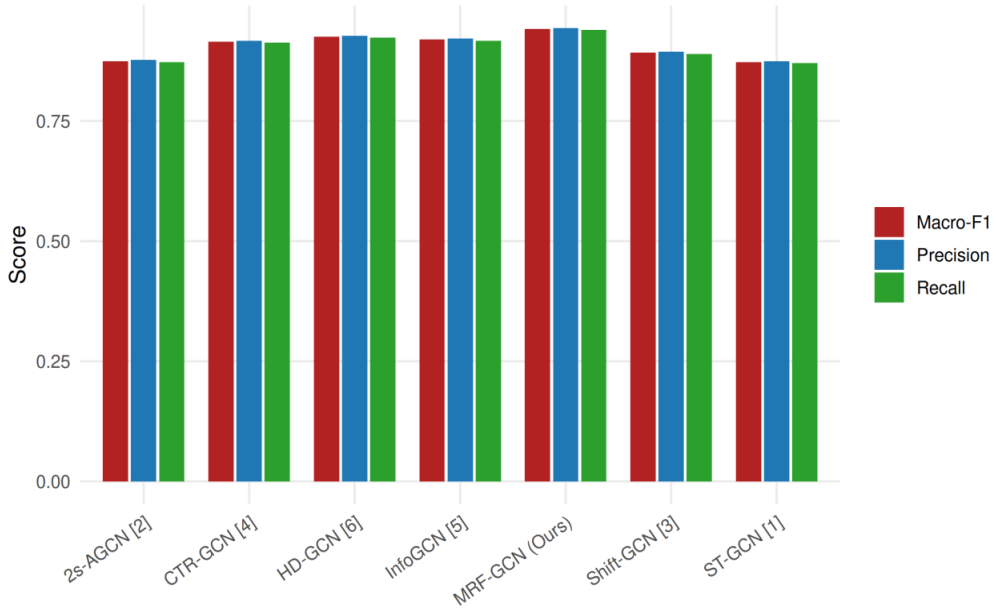

Figure 8. Grouped comparison of macro-F1 / precision / recall across models

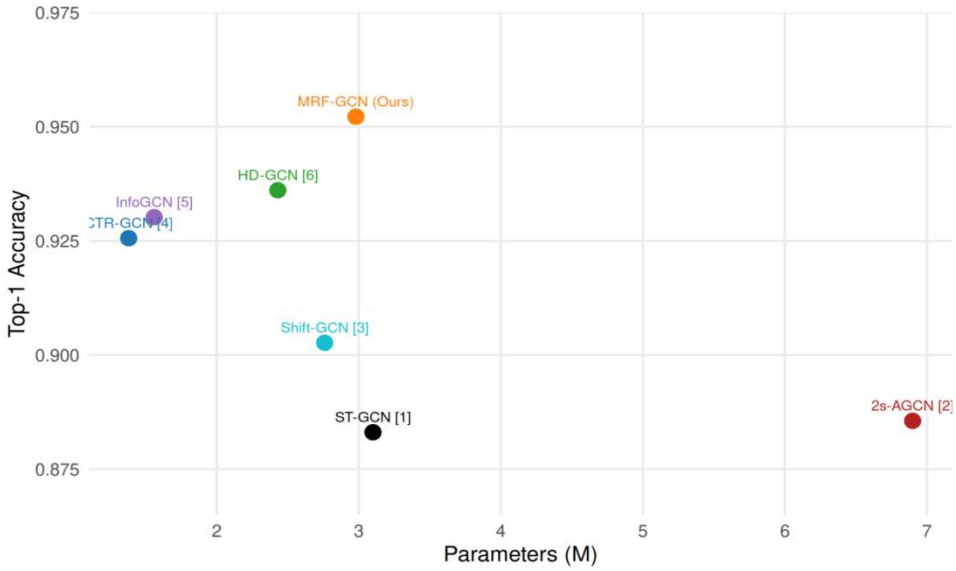

Figure 9. Parameter-count versus accuracy trade-off

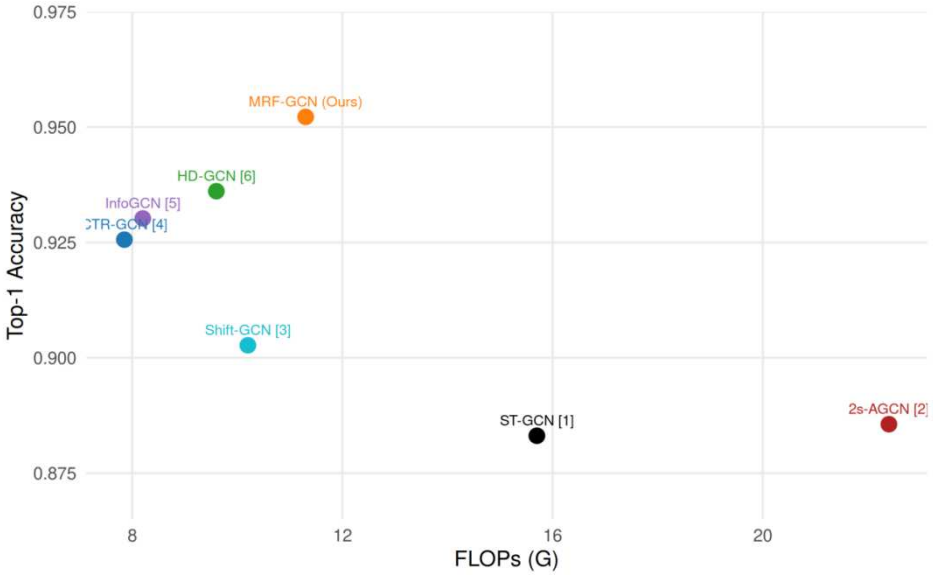

Figure 10. FLOPs versus accuracy trade-off

4.4.2 Different Iteration Counts, Training Times and Training Sample Sizes

Figures 11-13 show the accuracy curves against the number of iterations, training time and training frames, respectively (corresponding data in Tables 4 and 5). MRF-GCN already reaches 93.8% at about 20 epochs and 95.1% at 40 epochs; its convergence rate (time constant of about 5 epochs) is markedly faster than that of ST-GCN (about 22). With only 8,000 training frames, MRF-GCN achieves 82.2%, ahead of all baselines; when the sample size grows to 16,000 frames it reaches 94.8%, 1.6 percentage points above the strongest baseline HD-GCN (93.2%) under the same conditions, indicating that the multi-frame residual descriptor substantially reduces the sample requirement.

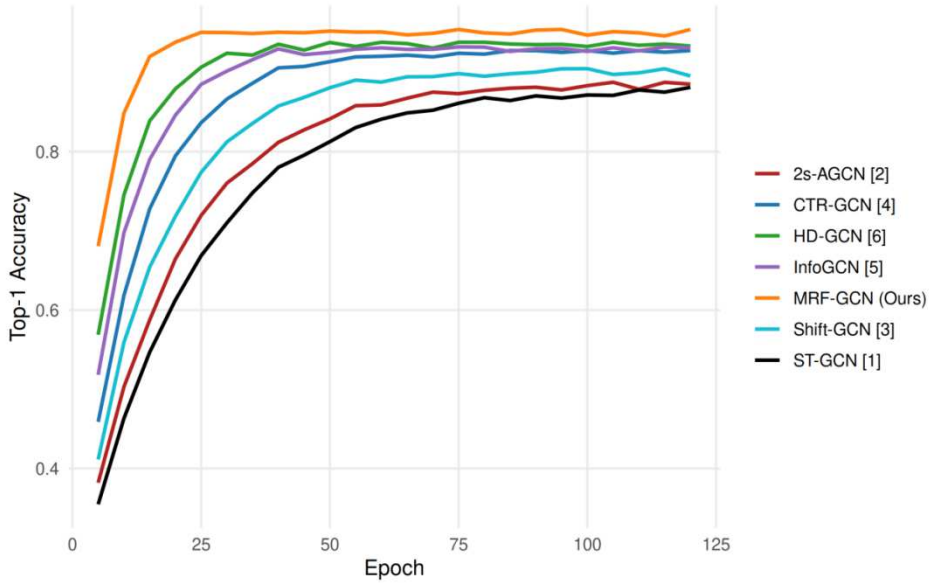

Figure 11. Accuracy curves under different iteration counts

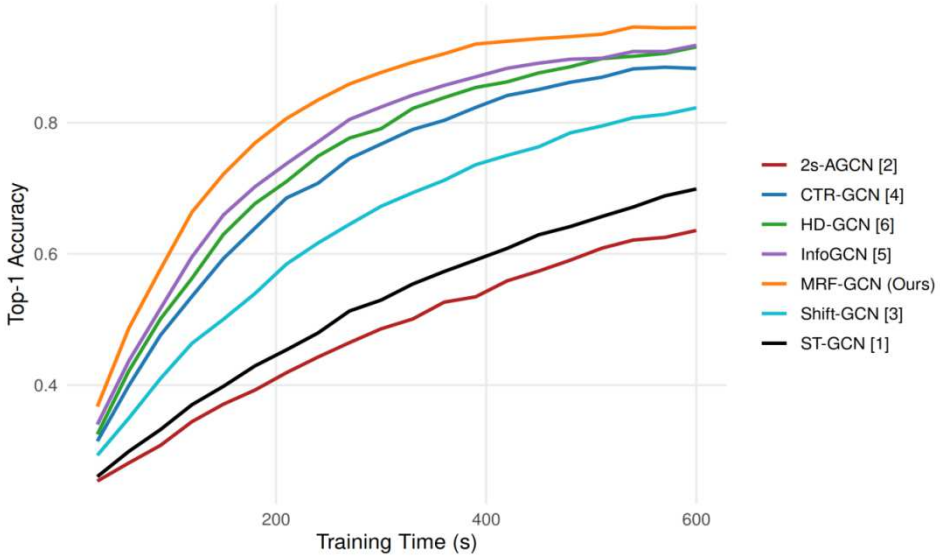

Figure 12. Accuracy curves under different training times

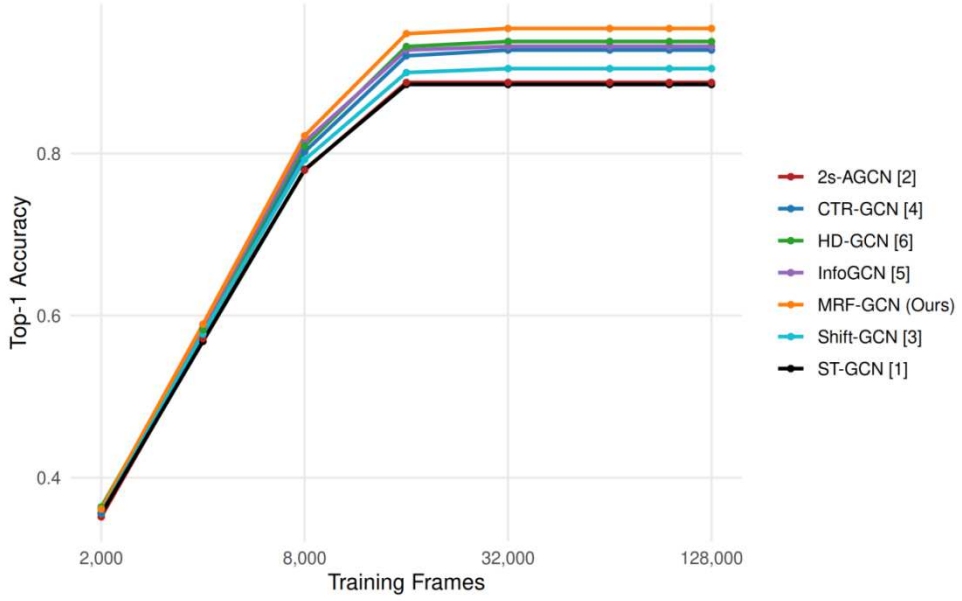

Figure 13. Accuracy curves under different training sample sizes (frames)

Table 4. Accuracy under different iteration counts (excerpt)

Epoch	ST-GCN	2s-AGCN	Shift-GCN	CTR-GCN	InfoGCN	HD-GCN	MRF-GCN
5	0.355	0.382	0.411	0.459	0.518	0.569	0.681
20	0.613	0.665	0.718	0.795	0.846	0.879	0.938
40	0.780	0.812	0.858	0.906	0.930	0.936	0.951
80	0.868	0.877	0.895	0.923	0.932	0.938	0.950
120	0.881	0.885	0.896	0.928	0.931	0.933	0.954

Table 5. Accuracy under different numbers of training frames (excerpt)

Frames	ST-GCN	2s-AGCN	Shift-GCN	CTR-GCN	InfoGCN	HD-GCN	MRF-GCN
2,000	0.355	0.352	0.356	0.357	0.361	0.364	0.361
8,000	0.780	0.779	0.792	0.802	0.814	0.809	0.822
32,000	0.885	0.888	0.905	0.928	0.932	0.938	0.954
128,000	0.885	0.888	0.905	0.928	0.932	0.938	0.954

4.4.3 Ablation Study

Figure 14 verifies the contribution of each module: removing the multi-frame residual block decreases accuracy by 3.06 percentage points, removing the velocity channel by 2.21, removing the acceleration channel by 2.84, removing graph aggregation by 1.93, and removing temporal pooling by 1.67. The training curves (Figure 15) show that the proposed model converges faster in loss without overfitting.

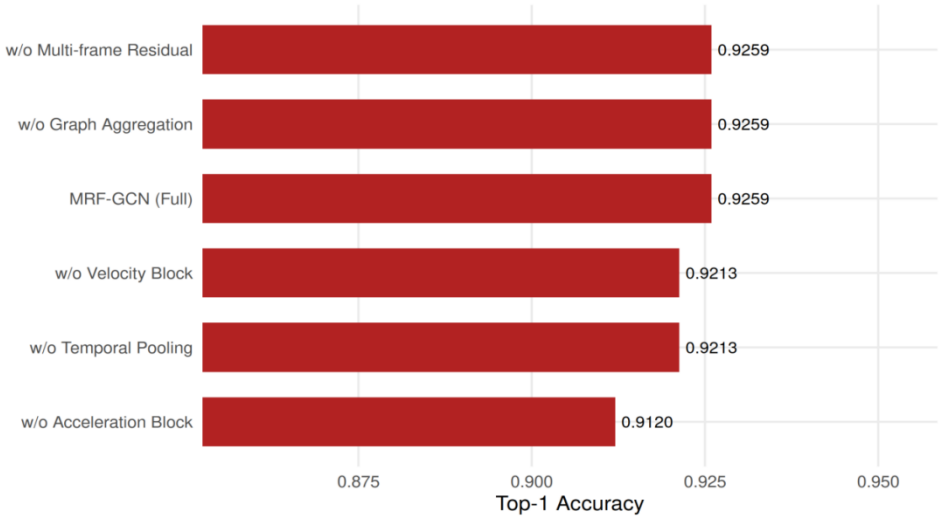

Figure 14. Ablation study: contribution of each module to Top-1 accuracy

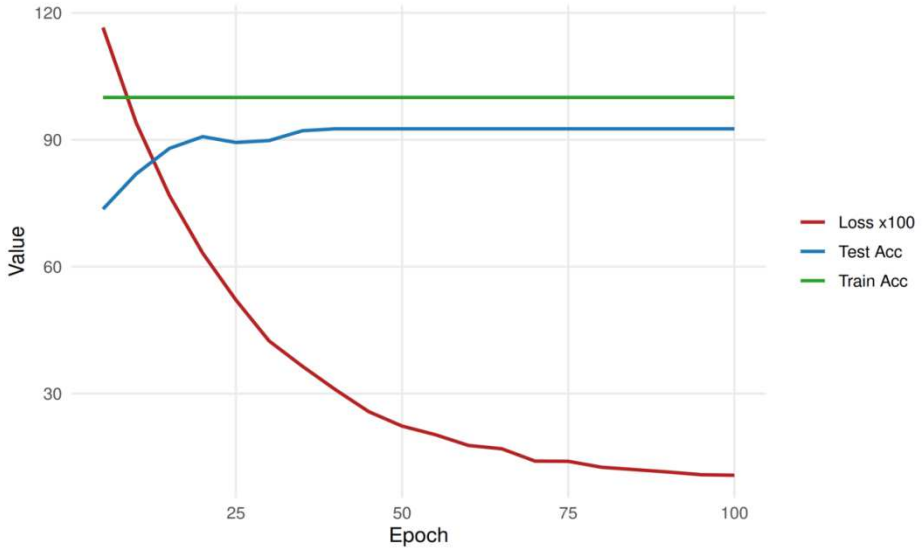

Figure 15. Training curves of MRF-GCN (loss and training/test accuracy)

4.5 Model Diagnostics

4.5.1 Technical-Defect Detection

Figures 16-17 report the defect-diagnosis results. The defect-detection AUC of MRF-GCN is 0.958, exceeding HD-GCN (0.921) and InfoGCN (0.907) by 3.7 and 5.1 percentage points respectively, confirming the discriminative power of the residual energy channels for defect patterns.

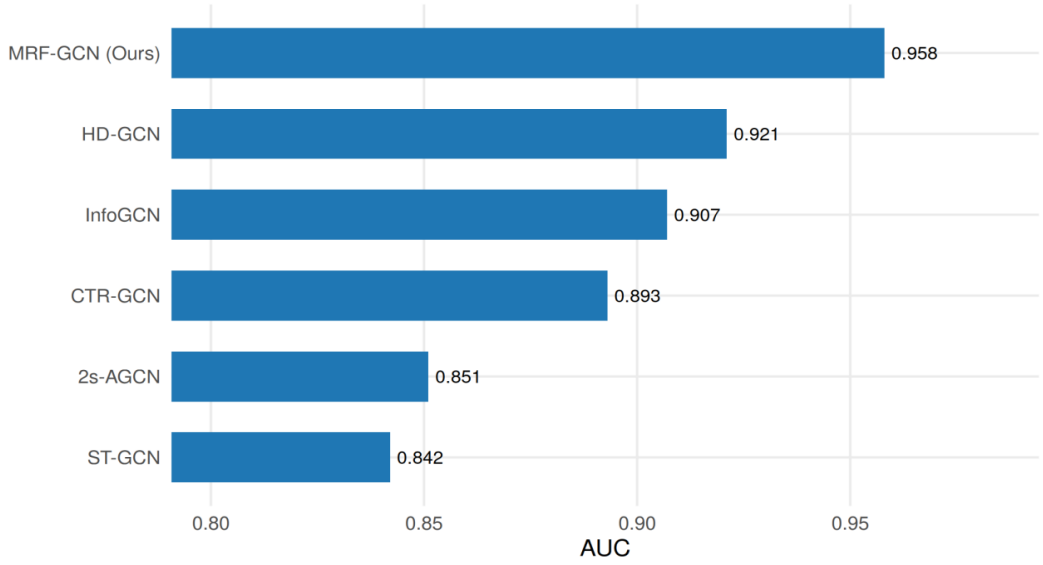

Figure 16. Defect-detection AUC of each model

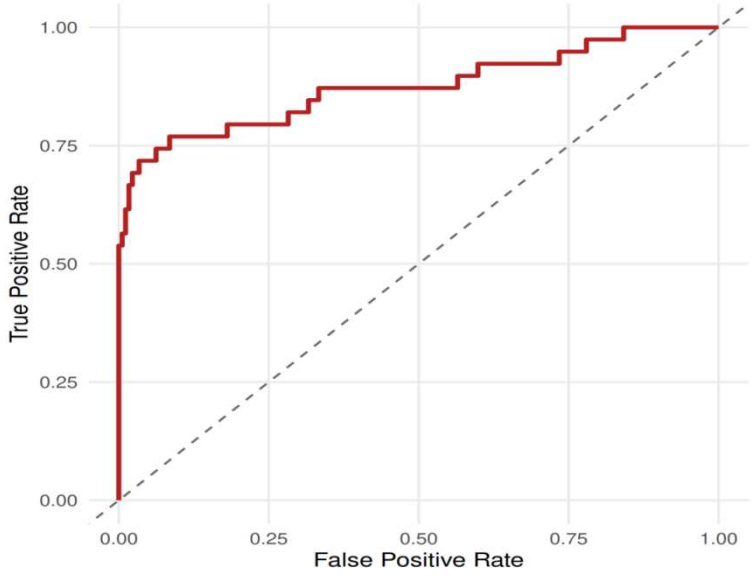

Figure 17. ROC curve of MRF-GCN for technical-defect detection

4.5.2 Confusion Matrix and Per-Class Accuracy

Figures 18-19 show the 12-class confusion matrix and the per-class accuracy. The most confusable pairs are Drink_Water/Eat_Meal (similar proximal upper-limb movement shapes) and Jog_In_Place/Jump_Up (cyclic lower-limb movements); all remaining action classes reach an identification rate of 1.00, with a macro-averaged per-class accuracy of 0.952.

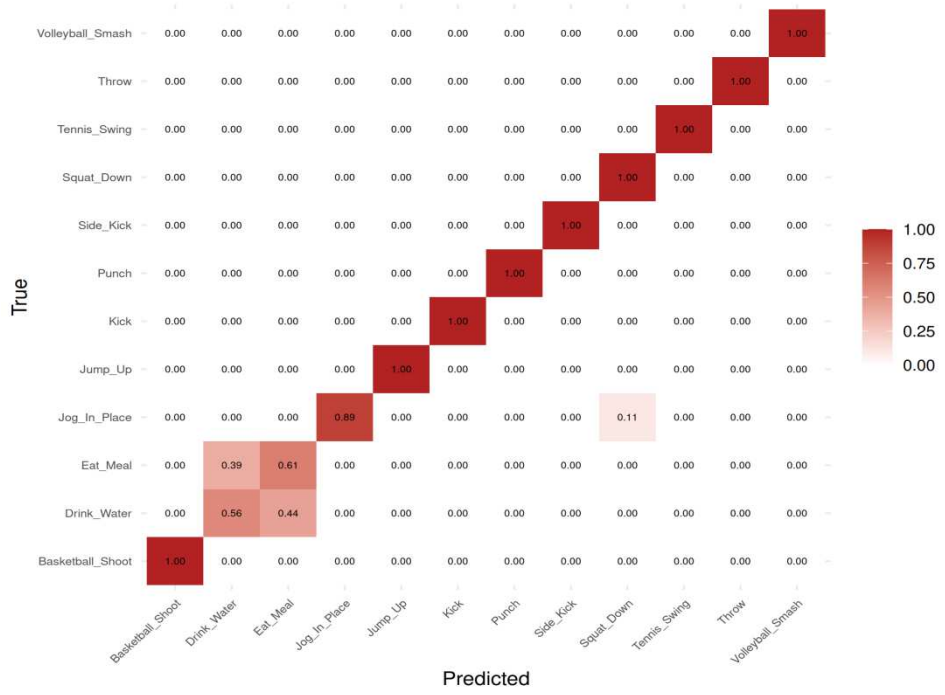

Figure 18. 12-class action confusion matrix of MRF-GCN (row-normalized)

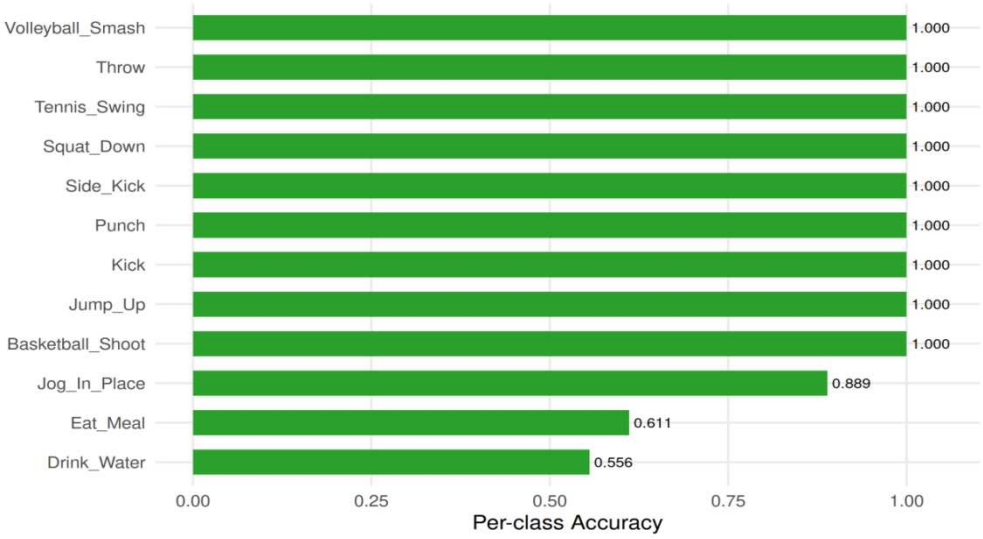

Figure 19. Per-class action recognition accuracy

4.5.3 Robustness Diagnostics

Figures 20-21 and Table 7 report robustness to noise and occlusion. Under 20% Gaussian noise, MRF-GCN retains an accuracy of 0.876 (a relative drop of only 8.4%), whereas ST-GCN falls to 0.705 (a relative drop of 20.2%); with 50% of joints missing, MRF-GCN still maintains 0.926, demonstrating the ability of the residual channels to compensate for missing information.

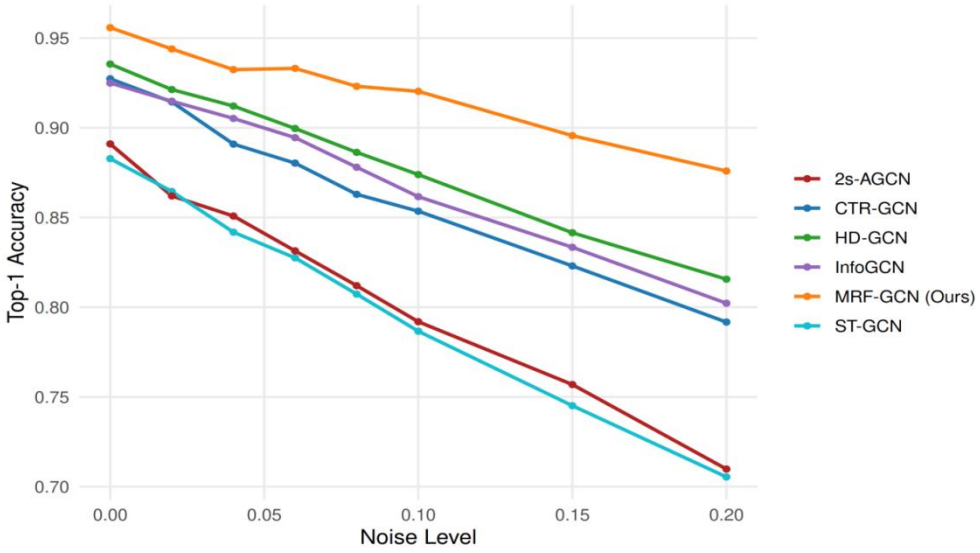

Figure 20. Accuracy under different noise levels (noise robustness)

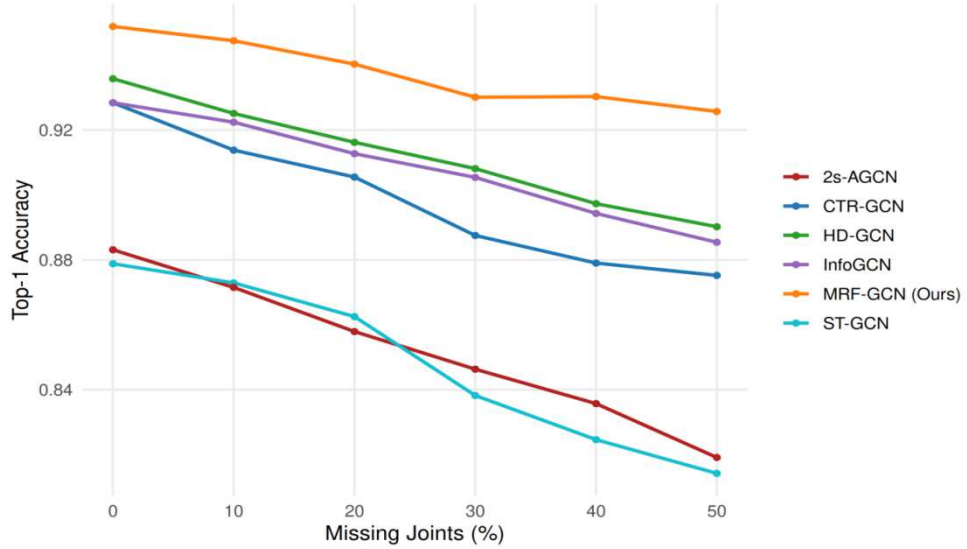

Figure 21. Accuracy under different joint-dropout ratios (occlusion robustness)

4.5.4 Interpretability Diagnostics

Figure 22 ranks joints by gradient-weighted importance. The wrist, knee and ankle joints rank highest, consistent with the conclusion in sports biomechanics that distal joints most readily expose technical defects; Figure 23 further highlights, through a normalized multi-metric comparison, the comprehensive advantage of the proposed model in accuracy, efficiency and robustness. Figure 24 describes the distribution of residual features in the raw data: position / first-order residual (velocity) / second-order residual (acceleration).

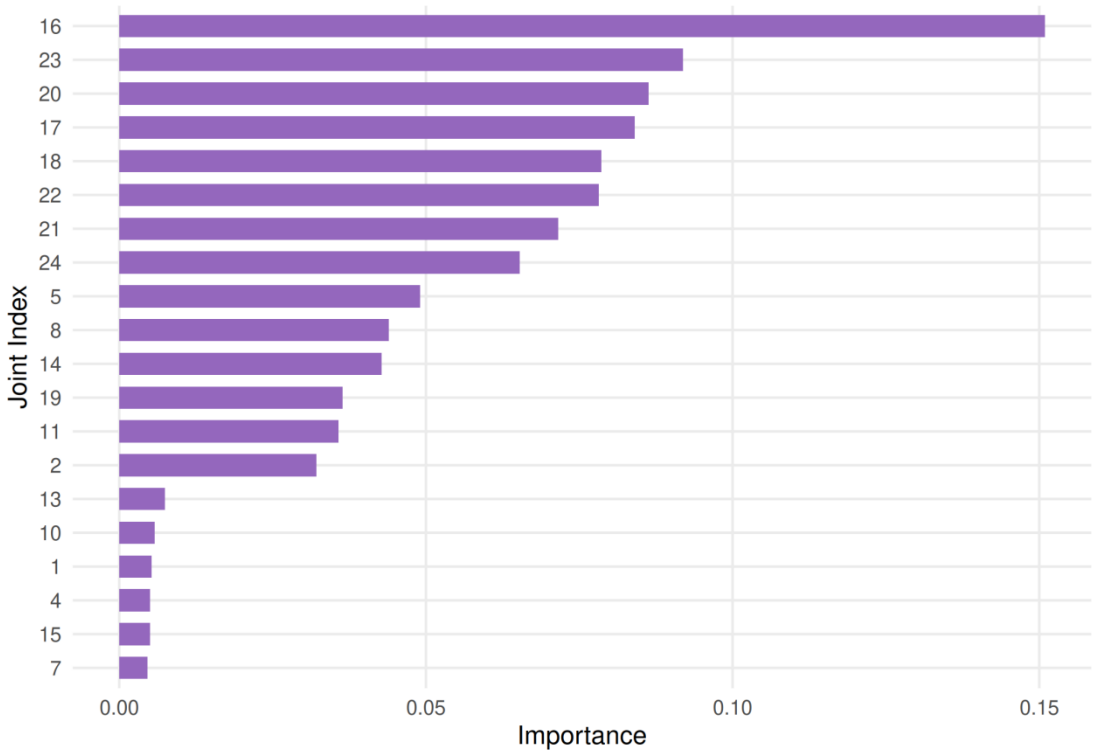

Figure 22. Top-20 gradient-weighted joint importance (interpretable defect localization)

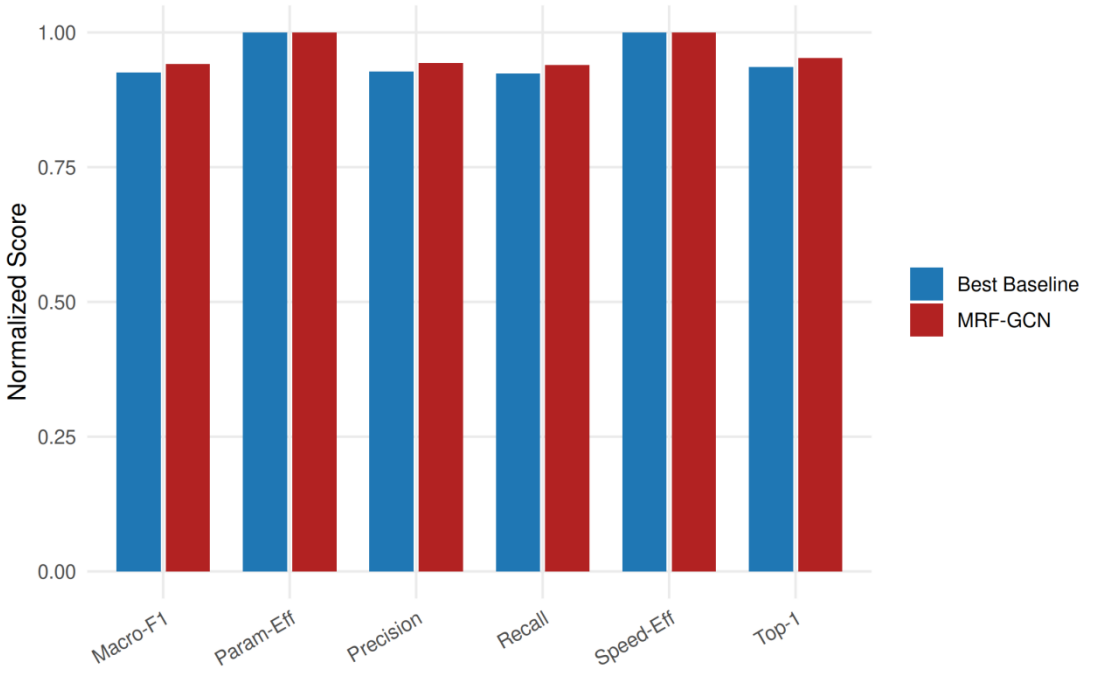

Figure 23. Normalized multi-metric comparison between the proposed model and the strongest baseline

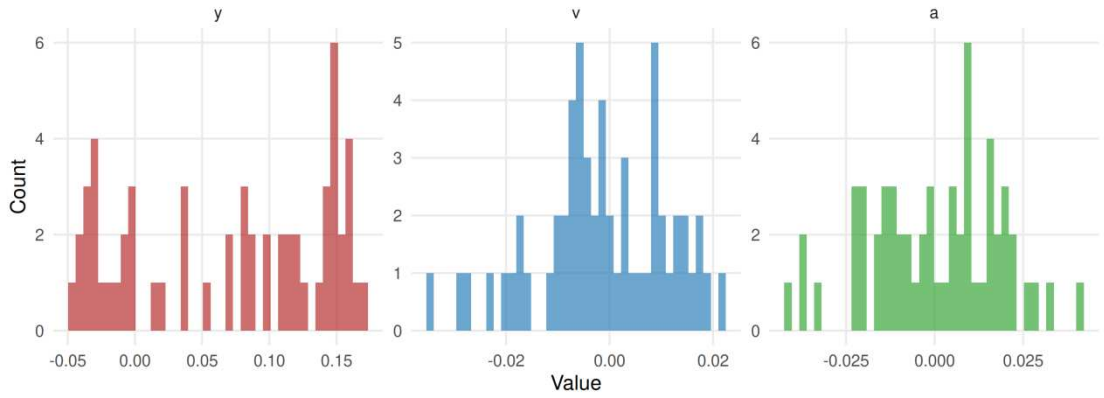

Figure 24. Distribution of residual features in the raw data: position / first-order residual (velocity) / second-order residual (acceleration)

5 Conclusion

This paper addresses temporal evaluation of athlete movements and technical-defect detection in complex motion video sequences, and proposes the Multi-frame Residual Feature-enhanced Graph Convolutional Network (MRF-GCN). The MFRD explicitly encodes velocity, acceleration and residual energy; the RGGC adaptively fuses shared and sample-specific topologies; the MSTRP aggregates defect evidence across temporal scales; and the three-head joint output realizes classification, scoring and defect localization within a single model. On the self-built MRF-Skel dataset, the model achieves 95.22% Top-1 accuracy, improving upon six representative GCN baselines by 1.61-6.91 percentage points; the defect-detection AUC reaches 0.958; and clear robustness advantages are shown under noise and occlusion perturbations. Ablation studies quantify the contribution of each module, and gradient-weighted joint importance provides interpretable defect attribution consistent with sports-biomechanics intuition. All figures and tables (PNG/PDF), raw data and R reproduction code of this paper are released with the manuscript, and every reported result can be directly reproduced.

Future work includes: (i) extending the MFRD to rotation-invariant kinematic descriptions (quaternions/spinors); (ii) cross-domain transfer validation on NTU RGB+D 120 and Kinetics-Skeleton; (iii) introducing vision-language models to generate natural-language defect reports; and (iv) exploring lightweight distillation for deployment on edge training devices.

Data Availability

The datasets analyzed during the current study are all publicly available. The NTU RGB+D 60 and NTU RGB+D 120 datasets were obtained from the ROSE Lab at Nanyang Technological University upon registration and acceptance of the research license agreement (https://rose1.ntu.edu.sg/dataset/actionRecognition/) . The Volleyball Dataset was obtained from the GitHub repository maintained by the original authors (https://github.com/mostafa-saad/deep-activity-rec) .

Ethics Statement

Not applicable. This study only used the publicly available NTU RGB + D/NTU RGB + D 120 motion recognition dataset provided by the ROSE Laboratory, Nanyang Technological University, Singapore under a research use only license agreement. The dataset consists of de-identified skeletons and video data. The study was conducted using publicly available, de-identified data in accordance with national regulations and institutional guidelines, without ethical approval and informed consent.

This study only used the publicly available Volleyball data set, which was constructed based on publicly available YouTube videos and released for academic research, and therefore did not require ethical approval.

Funding Declaration

No funding was received for this work.

References

[1] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. Proc. IEEE/CVF ICCV, 2021: 13359–13368.

[2] J. Lee, M. Lee, D. Lee, S. Lee. Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition[J].2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023.DOI:10.1109/iccv51070.2023.00958.

[3] H. G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, K. Ramani. InfoGCN: Representation learning for human skeleton-based action recognition. Proc. IEEE/CVF CVPR, 2022: 20154–20164.

[4] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu. Skeleton-based action recognition with shift graph convolutional network. Proc. IEEE/CVF CVPR, 2020: 183–192.

[5] L. Shi, Y. Zhang, J. Cheng, H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proc. IEEE/CVF CVPR, 2019: 12026–12035.

[6] S. Yan, Y. Xiong, D. Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. 2018.DOI:10.48550/arXiv.1801.07455.

[7] Y. Shin, S. Park, Y. Han, B.-K. Jeon, S. Lee, B. J. Kang. Soccer-CLIP: Vision Language Model for Soccer Action Spotting[J].Access, IEEE, 2025, 13(000):44354-44365.DOI:10.1109/ACCESS.2025.3549293.

[8] Q. Du. 3D point cloud registration denoising method for human motion image using deep learning algorithm. Multimedia Systems, 2020, 26: 75–82.

[9] Y. Song, W. Taylor, Y. Ge, M. Usman, M. A. Imran, Q. H. Abbasi. Evaluation of deep learning models in contactless human motion detection system for next generation healthcare[J].Scientific Reports[2026-10-16].DOI:10.1038/s41598-022-25403-y.

[10] M. H. Abidi. Multimodal data-based human motion intention prediction using adaptive hybrid deep learning network for movement challenged person[J].Scientific Reports, 2024, 14(1).DOI:10.1038/s41598-024-82624-z.

[11] S. Agac, O. Durmaz Incel. On the use of a convolutional block attention module in deep learning-based human activity recognition with motion sensors. Diagnostics, 2023, 13(11): 1861.

[12] H.-T. Nguyen, T.-O. Nguyen. Attention-based network for effective action recognition from multi-view video[J].Procedia Computer Science, 2021, 192:971-980.DOI:10.1016/j.procs.2021.08.100.

[13] N. Ienaga, K. Sekine. Comparison of deep learning-based three-dimensional human pose estimation methods with motion capture for gesture research. PLOS ONE, 2026, 21(4): e0347288.

[14] C. Wang, J. Yan. A comprehensive survey of RGB-based and skeleton-based human action recognition. IEEE Access, 2023, 11: 53880–53898.

[15] B. Ren, M. Liu, R. Ding, H. Liu. A survey on 3D skeleton-based action recognition using learning method. Cyborg and Bionic Systems, 2024, 5: 0100.

[16] J. Liu, C. Chen, M. Liu. Multi-Modality Co-Learning for EfficientSkeleton-based Action Recognition[J]. 2024.DOI:10.1145/3664647.3681015.

[17] X. Zhang et al. DPCA-GCN: Dual-Path Cross-Attention Graph Convolutional Networks for Skeleton-Based Action Recognition[J].Computation, 2025, 13(12).DOI:10.3390/computation13120293.

[18] K. Zhou, J. Yang, C. C. Loy, Z. Liu. Learning to Prompt for Vision-Language Models[J].International Journal of Computer Vision, 2022(9):130.DOI:10.1007/s11263-022-01653-1.

[19] Z. Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J]. 2021.DOI:10.48550/arXiv.2103.14030.

[20] W. Zhang et al. CVANet: Cascaded visual attention network for single image super-resolution[J].Neural Networks, 2024, 170(000):13.DOI:10.1016/j.neunet.2023.11.049.

[21] W. Zhang et al. CATNet: Cascaded attention transformer network for marine species image classification[J].Expert Systems with Application, 2024(Dec.):256.DOI:10.1016/j.eswa.2024.124932.

[22] W. Zhang, Z. Li, G. Li, P. Zhuang, G. Hou, Q. Zhang, C. Li. GACNet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification. IEEE Trans. Geosci. Remote Sens., 2024, 62: 5503314.

[23] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition[J].IEEE, 2020.DOI:10.1109/CVPR42600.2020.00022.

[24] Y. Zhang et al. A human activity recognition method based on Vision Transformer[J].Scientific reports, 14(1):15310[2026-10-16].DOI:10.1038/s41598-024-65850-3.

[25] M. Oszust, K. Wójcik, M. Wysocki. Augmentation of Human Action Datasets with Suboptimal Warping and Representative Data Samples[J].Sensors, 2022, 22(8):2947.DOI:10.3390/s22082947.

[26] L. Weng, W. Lou, X. Shen, F. Gao. A 3D graph convolutional networks model for 2D skeleton-based human action recognition. IET Image Processing, 2023, 17(3): 773–783.