Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STEP-FORMER: RELIABILITY-AWARE SPATIO-TEMPORAL POSE MODELING FOR LIGHTWEIGHT HUMAN ACTION RECOGNITION UNDER IMPERFECT POSE ESTIMATES

B. PANDU RANGA RAJU[1], DR. CH. VENKATA RAMANA REDDY[2], DR. K. PRADEEP REDDY[3]

[1]Research Scholar, Department of CSE, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha, India.
[2]Professor, Department of CSE, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha, India.
[3]Professor, CSE Department, Vignan's Institute of Management and Technology for Women, Hyderabad, Telangana, India.
Corresponding author E-mail: bpanduranga.raju@giet.edu

ABSTRACT

Human action recognition using pose decreases the importance of appearance cues and allows for privacy-aware recognition; however, recognition accuracy may decrease due to missing, jittered, or inconsistently detected two-dimensional poses. Recent graph-, transformer-, temporal-pooling-, hypergraph-, and language-assisted skeleton approaches are commonly evaluated with different skeleton dimensionalities, modalities, datasets, or preprocessing pipelines; consequently, their published scores are not directly comparable with a controlled 2D OpenPose setting. In this work, we propose STeP-Former, a lightweight pose-only recognizer that uses Spatio-Temporal Pose Abstraction Layer (ST-PAL), label-free Reliability-Aware Temporal Fusion (RTF), and shallow temporal attention-based encoder. ST-PAL maps normalized joint coordinates and their first order derivatives to motion tokens. RTF computes the reliability of individual frames from the consistency of local tokens and penalizes unreliable frames before global temporal modeling. The evaluation is carried out following the usual disjoint protocol for KTH dataset and the official UCF101 Split 1. Four baseline models that use poses are reproduced on the same cached OpenPose BODY-25 sequences with three fixed seeds, while RGB-based or protocol non-compatible recent approaches are provided separately for reference purposes as literature only. STeP-Former achieves 95.6 ± 0.28% accuracy on KTH and 90.2 ± 0.41% accuracy with 0.897 ± 0.007 macro-F1 score on UCF101. The recognizer has 2.1 M parameters and takes 9 ms to process 32 frames on a clip according to the mentioned recognizer-only timing protocol. Under the two controlled empirical stress conditions, STeP-Former shows smaller accuracy degradation than the aggregate pose baselines, while the simulation-only results are retained as sensitivity analyses rather than benchmark evidence.

Keywords: *Human Action Recognition, Two-Dimensional Pose Sequences, Motion Tokens, Reliability-Aware Temporal Fusion, Temporal Attention, Recognizer Efficiency*

HIGHLIGHTS

1. ST-PAL tokens jointly embed normalised pose and first-order joint motion without any appearance information from RGB.
2. The RTF method utilises a label-free temporal consistency gate before self-attention to decrease local inconsistency of poses.
3. Pose baseline controlled rerun, literature-only experiments, ablation study, and latency estimation make our work fair and reproducible.

1. INTRODUCTION

HAR applications include surveillance, assistive living, rehabilitation, sports analysis, and human-computer interaction. Three-dimensional convolutions for RGB data and video transformers are capable of learning robust visual-motion representations; however, their computational expense and reliance on scene context can limit their deployment to privacy-concerned or resource-strapped scenarios [3], [4], [8]. The lighting, viewing angle, background, and recording setup can also change the appearance statistics of an action without affecting its essence.

The pose-based approach to HAR utilizes trajectories of body joints rather than dense contents of images. The advantage of this kind of data is that it requires less visual information for

recognition; however, the accuracy of pose estimation is important for this approach. Motion blur, self-occlusion, truncation, changes in scale, and presence of multiple people can cause errors such as missing joints, abrupt coordinate shifts, and even identity change. Temporal uniformity makes it possible for a few bad frames to affect the whole sequence. Thus, the goal of this research is not to achieve robustness across all conditions but rather to explore whether a compact two-dimensional pose recognizer is able to decrease sensitivity to intermittent pose inconsistencies without increasing the complexity of recognition. STeP-Former achieves this goal by encoding both posture and velocity into motion tokens, evaluating frame reliability based on temporal consistency, and using shallow temporal attention encoder only for filtered tokens. Three design tenets guide our work:

(i) normalise the pose geometry so that the recogniser is invariant to the image resolution;

(ii) learn reliability from the posed sequence itself without using extra confidence labels; and

(iii) allocate computational resources to temporal inference on a compact token sequence rather than spatial analysis.

This yields contributions that are analysed using an experimental protocol that is pose-centric, with RGB or three-dimensional skeleton protocols yielding incompatible results excluded from consideration.

1.1 Motivation and Novelty

All four components of STeP-Former—joint positions, time difference, reliability weight, and self-attention—have some relation to existing concepts in skeleton recognition. The difference in approach comes in the order and extent to which the components are used: a pose-velocity token is generated in a small two-dimensional format first; the reliability weight, based on residual values between tokens locally, is calculated second and applied prior to the temporal self-attention mechanism; and neither pose-confidence, RGB data, language description, nor reliability annotations are provided to the classifier. Therefore, the paper introduces STeP-Former as a reliability-based framework for imperfect two-dimensional poses and not an alternative to graph-, hyper-graph-, or multimodal skeleton models.

1.2 Research Questions and Objectives

To make the research problem testable and to connect the proposed design directly to the evaluation, the study is organized around the following research questions and corresponding objectives.

RQ1. Under an identical two-dimensional OpenPose BODY-25 protocol, can compact pose-velocity tokenization combined with reliability-aware temporal fusion improve recognition accuracy and macro-F1 over controlled pose-based baselines?

RQ2. Does applying a label-free reliability gate before temporal self-attention reduce performance degradation when intermittent pose corruption or viewpoint-related body truncation is present?

RQ3. Can the above robustness be obtained with a lightweight recognizer whose model size and recognizer-only latency remain suitable for resource-constrained deployment studies?

Objective O1. Design a compact two-dimensional pose recognizer that jointly represents normalized posture and first-order motion and suppresses locally inconsistent frames before long-range temporal aggregation.

Objective O2. Evaluate recognition, component contribution, and robustness using fixed KTH and UCF101 protocols, shared cached poses, controlled baseline reruns, ablation, and clearly separated empirical versus simulation-based diagnostics.

Objective O3. Quantify recognizer complexity and latency while explicitly separating recognizer timing from video decoding, pose estimation, tracking, and preprocessing costs.

1.3 Key Contributions

• An efficient ST-PAL representation is derived using normalized BODY-25 coordinates and velocity vectors, allowing modeling of both posture and motion within a unified token space.

• A label-independent RTF gate is designed prior to temporal self-attention and uses local token consistency, but does not utilize the OpenPose confidence values as classification features.

• A fair evaluation boundary is defined: ShiftGCN++, 2s-AGCN, TranSkeleton, and DeGCN are re-evaluated using the same pose cache and seed settings, and the recently proposed methods that work solely in the RGB modality or have protocol mismatches are recognized as literature-only baselines.

• Component ablation study shows that the motion tokens and reliability fusion modules add independent contributions; experiments with controlled occlusions and truncations define the robustness evidence more strictly.

• STeP-Former yields 95.6 ± 0.28% KTH accuracy and 90.2 ± 0.41% UCF101 accuracy using 0.897 ± 0.007 macro-F1, 2.1 M recognizer parameters, and 9 ms recognizer inference time per 32-frame video segment.

1.4 Organization of the Paper

The remaining sections are structured as follows. Section 2 critically reviews video-, skeleton-, reliability-, and robustness-oriented HAR and defines the fair-comparison boundary. Section 3 formulates the imperfect-pose recognition problem, and Section 4 presents STeP-Former. Section 5 describes the complete experimental protocol, including datasets, pose caching, baseline harmonization, training, evaluation, robustness analysis, and timing. Section 6 reports recognition, efficiency, ablation, controlled stress tests, simulations, and diagnostics. Section 7 interprets the findings and explicitly distinguishes the present contribution from prior research; Section 8 summarizes the study scope, principal limitation, and future directions; and Section 9 answers the research questions and summarizes the extent to which the objectives were achieved.

2. RELATED WORK

2.1 Video-centric Human Action Recognition

Video-based HAR moved from handcrafted features to two-stream and 3D convnets, followed by recent advances in video transformer networks [3], [4], [8], [11]. The above models learn appearances and motion together and can perform effectively in large-scale video benchmark tests. But the dense spatial computation makes these models memory-intensive and slow, and their predictions can depend on contextual information such as background or object. In the current study, the RGB approaches are considered as references to literature only, while not as unified protocol-based baselines for pose-only recognition. Other related work on video includes event camera-based action recognition, early prediction of egocentric actions, semantics-based video models, action detection in untrimmed videos, local-global representations, and domain-specific sports recognition [5], [6], [13], [21], [24], [29].

2.2 Pose-centric and Skeleton-based HAR

Skeleton-based HAR relies on recurrent neural networks, graph convolutional neural networks (GCNs), temporal convolution, pooling, and transformer architectures [2], [12], [16], [18], [19], [20]. Some recent developments include deformable graph topology [16], multi-view time series hyper-graph [14], self-attention GCN [25], asynchronous joint-based temporal pooling [7], skeleton-language supervision [9], and dynamic graph partitioning [31]. These approaches enhance either the spatial graph topology, temporal sampling or semantics supervision; however, most of these techniques have been tested against three-dimensional skeleton benchmarks like NTU RGB+D or PKU-MMD and their performance figures are not comparable to our OpenPose evaluation on KTH and UCF101 because of their different settings. Other possible graph configurations are adaptive temporal modelling [1], central difference operator [15], contrastive two streams [17], perceptually enhanced graphs [23], feedback graphs [27], scene context aware graphs [28], and spatial adaptive graphs [30]. A survey of Kinect action recognition approaches can be found in [22]; temporal action segmentation [10] and unsupervised skeleton representation learning [26] tackle related, but slightly different problems.

2.3 Reliability, Uncertainty, and Robust Aggregation for Pose Sequences

Pose-informed skeleton recognition aims at lowering the number of errors due to absence or noise in the joint locations. State-of-the-art methods incorporate confidence weighting, uncertainty quantification, topology adaptation, augmentation, or robust pooling. Pose-error augmentation has been proven to be beneficial for skeleton recognition under erroneous pose estimation [35]. However, confidence-weighting relies on specific score values from an estimator, augmentation adds choices during training, and topology-adaptive graphs become more complex. STeP-Former defines scalar weights for frames based on token consistency and then uses them prior to temporal attention. It is a deliberate design choice made not because all pose errors can be fixed with it.

2.4 Positioning and Fair-Comparison Scope

Table 1 highlights the comparison of STeP-Former with representative methods starting from 2024. The experiments included only those methods which are re-implemented using the same BODY-25 cache by OpenPose. The methods involving three-dimensional skeletons, language description, RGB features, other datasets, or inaccessible preprocessing are mentioned either in qualitative comparison or literature-only tables. This approach helps to avoid confusion while presenting the results obtained by protocols other than a unified experiment.

Table 1. Recent skeleton-action methods and fair-comparison boundary

Method	Primary contribution, limitation relative to the present problem, and comparison status
DeGCN [16] (2024)	Deformable graph topology improves spatial relation modeling, but it does not explicitly estimate frame reliability from intermittent 2D pose inconsistency; rerun on BODY-25 for the controlled UCF101 comparison.

MV-TS Hypergraph [14] (2024)	Higher-order multi-view temporal relations capture complex dependencies, but published results use different 3D skeleton protocols; therefore used only as literature context.
SelfGCN [25] (2024)	Self-attention graph convolution strengthens graph feature learning, but it targets graph representation rather than label-free frame reliability and uses a different benchmark/skeleton representation.
AJTP [7] (2025)	Asynchronous joint-based temporal pooling selects informative joint-time regions, but it addresses temporal selection rather than pre-attention reliability weighting and is reported under a different protocol.
Skeleton-language [9] (2025)	Language-supervised skeleton features add semantic guidance, but the additional modality and pretraining assumptions prevent a direct pose-only comparison.
DPGCN [31] (2026; online 2025)	Dynamic graph partitioning adapts spatial structure, but it was not reimplemented on the present 2D BODY-25 cache and does not isolate the same intermittent-pose reliability question.
STeP-Former	Pose-velocity tokens plus a label-free pre-attention reliability gate; evaluated using the controlled 2D BODY-25 protocol. Its scope is deliberately limited to the tested pose pipeline rather than a cross-protocol state-of-the-art claim.

2.5 Critical Synthesis, Research Gap, and Need for the Present Study

Recent skeleton-HAR research improves different parts of the recognition pipeline, but the improvements do not address the same failure mechanism. Topology-focused methods such as DeGCN [16], SelfGCN [25], and DPGCN [31] strengthen spatial graph construction or partitioning; MV-TS Hypergraph [14] models higher-order relations; and AJTP [7] selects informative joint-time regions. These approaches are valuable for representation learning, yet they are primarily evaluated on different skeleton dimensionalities or benchmark protocols and do not explicitly ask whether a small number of unreliable frames from a two-dimensional pose estimator should be down-weighted before global temporal attention.

Robustness-oriented studies are more closely related to the present problem. Cormier et al. [35] improve real-world skeleton recognition through realistic pose-error augmentation, while Wu et al. [36] address missing keypoints with an explicit two-stage interpolation and masked-autoencoder pipeline. Skeleton-Prompt [37] improves cross-dataset transfer and occlusion robustness through pretraining, prompting, and multi-stream fusion. Their strengths also define a different design cost: robustness is obtained through corruption-aware augmentation, reconstruction, pretraining, or additional streams. By contrast, STeP-Former tests whether reliability can be inferred directly from the internal temporal consistency of the observed 2D pose tokens, without corruption labels, pose-confidence features, RGB appearance, language supervision, a pose-reconstruction network, or cross-dataset pretraining.

The resulting gap is therefore not a claim that graph, hypergraph, reconstruction, augmentation, or transfer-learning methods are ineffective. The unresolved IT question is narrower: whether a compact 2D pose recognizer can reduce sensitivity to intermittent estimator errors through a label-free pre-attention reliability signal, while preserving a small recognizer and a reproducible pose-centric comparison protocol. Across the recent studies reviewed in Table 1 and Refs. [35]–[37], the dominant contributions fall into spatial-topology refinement, higher-order temporal modeling, training-time corruption augmentation, explicit missing-keypoint reconstruction, semantic supervision, or cross-dataset transfer. The present work is positioned differently because it tests frame reliability directly from local token consistency before global temporal attention and evaluates that mechanism on a single cached BODY-25 pipeline with controlled reruns, ablation, stress testing, and recognizer-only efficiency measurement.

This positioning also defines the significance and the boundary of the contribution. A favorable result would show that a lightweight reliability gate can complement, rather than replace, graph-topology, augmentation, reconstruction, or transfer-learning strategies. Conversely, the present design should not be interpreted as superior to recent methods whose gains arise from different data modalities, 3D skeleton definitions, pretraining resources, or benchmark protocols. This distinction is maintained throughout the Results and Discussion sections.

3. PROBLEM FORMULATION

Consider a clip with T frames. OpenPose extracts $J = 25$ two-dimensional joints from each frame. After subject selection and missing-joint handling, the normalized pose at time t is represented by a vector $\boldsymbol{p}_t \in \mathbb{R}^{2J}$. OpenPose confidence values are used only to select the primary person and identify missing joints; they are not supplied to STeP-Former.

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\boldsymbol{p}_t = \left[x_1^t, y_1^t, \ldots, x_J^t, y_J^t\right]^\top \in \mathbb{R}^{2J}. \quad (1)$$

The complete normalized pose sequence is denoted by $\boldsymbol{P} = \{\boldsymbol{p}_t\}_{t=1}^{T}$.

$$\boldsymbol{P} = \{\boldsymbol{p}_t\}_{t=1}^{T}. \quad (2)$$

Short-term motion is represented by the first-order difference $\Delta\boldsymbol{p}_t = \boldsymbol{p}_t - \boldsymbol{p}_{t-1}$, with $\Delta\boldsymbol{p}_1$ defined as the zero vector. This finite-difference term preserves the direction and relative magnitude of local joint motion without introducing optical flow or RGB features.

$$\Delta\boldsymbol{p}_t = \boldsymbol{p}_t - \boldsymbol{p}_{t-1}, t = 2, \ldots, T; \Delta\boldsymbol{p}_1 = 0. \quad (3)$$

Let $y \in \{1, \ldots, C\}$ be the ground-truth action label for C classes. The model learns a mapping from the pose sequence to a class-probability vector.

$$f_\theta: \boldsymbol{P} \mapsto \hat{\boldsymbol{y}}, \hat{\boldsymbol{y}} \in [0,1]^C. \quad (4)$$

The parameters are optimized using categorical cross-entropy over the labeled training set. Accuracy and macro-F1 are reported on the untouched official test partitions.

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1\{y^{(i)} = c\} log\hat{y}_c^{(i)}. \quad (5)$$

The problem is one of intermittent corruptions and not consistent bad poses in every single frame. The problem may be local within the frame even if the neighbouring frames are consistent. In the case of STeP-Former, this weight is computed prior to the long-range aggregation process. This evaluation will check the performance of this mechanism using existing benchmarks and two perturbation cases.

4. PROPOSED METHOD: STEP-FORMER

The STeP-Former model normalises joint trajectories to motion tokens, reduces inconsistency among locally incompatible tokens, and uses a shallow transformer model for modelling the rest of the sequence. The computation moves from the processing of pixel-based features to temporal modelling of T = 32 pose tokens. Figure 1 presents the whole recognition pipeline and the divide between the external pose estimator and the trainable recogniser.

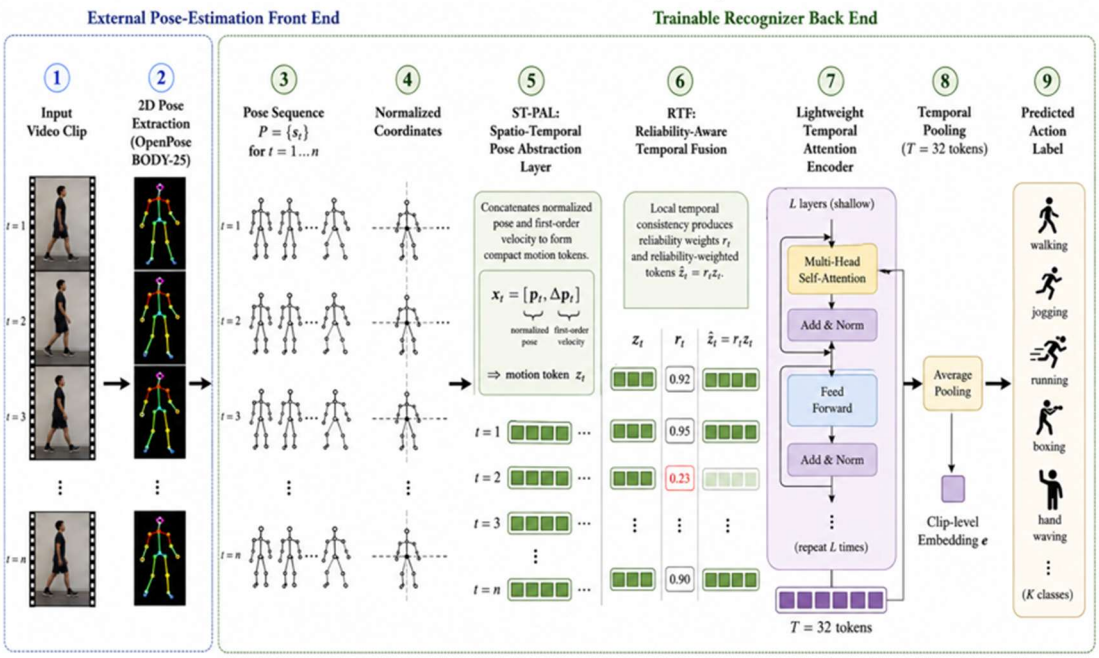

Figure 1. Overall architecture of STeP-Former from video frames and OpenPose BODY-25 extraction to ST-PAL, RTF, temporal attention, pooling, and action prediction.

4.1 Spatio-Temporal Pose Abstraction Layer (ST-PAL)

For each frame, the BODY-25 coordinates are centered and scaled using the valid torso joints. When a torso scale cannot be computed, the nearest valid scale in the sequence is used. ST-PAL concatenates the normalized pose $\boldsymbol{p}_t$ with its first-order difference $\Delta\boldsymbol{p}_t$ to form a motion descriptor $\boldsymbol{m}_t$. This representation allows actions with similar static configurations but different velocity patterns to be separated.

$$\boldsymbol{m}_t = [\boldsymbol{p}_t^\top, \Delta\boldsymbol{p}_t^\top]^\top \in \mathbb{R}^{4J}. \quad (6)$$

The motion vector $\boldsymbol{m}_t$ is mapped into a compact token space using a learnable projection followed by the non-linear activation $\phi(\cdot)$.

$$\boldsymbol{z}_t = \phi(\boldsymbol{W}_a \boldsymbol{m}_t + \boldsymbol{b}_a), \boldsymbol{z}_t \in \mathbb{R}^d. \quad (7)$$

The descriptor is projected to a d = 128 token through a learnable affine layer followed by GELU activation. Because the token contains only normalized pose and velocity, its dimension does not depend on image resolution or background content. The first frame uses a zero velocity vector, and the same tokenization is applied to every baseline that accepts vectorized pose input.

4.2 Reliability-Aware Temporal Fusion (RTF)

RTF assigns a scalar reliability weight to each token from its local temporal deviation. The score compares $\boldsymbol{z}_t$ with the mean neighboring token $\overline{\boldsymbol{z}}_t$, computed from the available immediate temporal neighbors; a boundary frame uses its single available neighbor. The fixed temperature $\tau > 0$ controls sensitivity to abrupt token changes.

$$r_t = exp\left(-\frac{\|\boldsymbol{z}_t - \overline{\boldsymbol{z}}_t\|_2^2}{\tau}\right), 0 < r_t \leq 1. \quad (8)$$

The reliability-weighted token is then obtained by

$$\tilde{\boldsymbol{z}}_t = r_t \boldsymbol{z}_t. \quad (9)$$

The temperature parameter τ controls how aggressively temporal outliers are suppressed and is fixed for all experiments. The operation is differentiable and introduces no learned reliability network. Importantly, OpenPose confidence is not included in the score; confidence is used only during preprocessing. RTF therefore tests whether sequence consistency alone is sufficient to reduce the influence of unstable frames.

4.3 Lightweight Temporal Attention Encoder

The reliability-weighted sequence is processed by $L = 2$ transformer encoder layers with four attention heads. Query, key, and value projections $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are computed from the same token sequence, and scaled dot-product attention models long-range dependencies among action phases. The symbol d_k denotes the key dimension of each attention head. Dropout is applied after the attention and feed-forward sublayers.

$$Attn(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}. \quad (10)$$

Self-attention has $\mathcal{O}(T^2 d)$ temporal cost. With $T = 32$ and $d = 128$, the temporal matrix is small compared with the number of patches processed by RGB video transformers. This comparison concerns the recognizer only; the external pose-estimation cost is reported separately as outside the timing scope of this study.

4.4 Temporal Pooling and Classification

The encoder returns contextualized tokens $\{\boldsymbol{h}_t\}_{t=1}^T$. A clip representation is obtained by temporal average pooling. The use of average pooling is retained to avoid adding a learned class token or another attention stage.

$$\boldsymbol{h} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t. \quad (11)$$

A linear classifier and softmax produce the final class-probability vector.

$$\hat{\boldsymbol{y}} = softmax(\boldsymbol{W}_c \boldsymbol{h} + \boldsymbol{b}_c). \quad (12)$$

Algorithm 1: STeP-Former Inference

Input: normalized pose sequence $\boldsymbol{P} = \{\boldsymbol{p}_t\}_{t=1}^T$

1. Set $\Delta\boldsymbol{p}_1 = 0$ and compute $\Delta\boldsymbol{p}_t = \boldsymbol{p}_t - \boldsymbol{p}_{t-1}$ for $t = 2, \ldots, T$.
2. Form $\boldsymbol{m}_t = [\boldsymbol{p}_t^\top, \Delta\boldsymbol{p}_t^\top]^\top$.
3. Project $\boldsymbol{m}_t$ to the d-dimensional ST-PAL token $\boldsymbol{z}_t$.
4. Compute a local temporal-consistency score r_t for every token.
5. Form reliability-weighted tokens $\tilde{\boldsymbol{z}}_t = r_t \boldsymbol{z}_t$.
6. Apply the two-layer temporal attention encoder to obtain $\boldsymbol{h}_t$.
7. Average-pool $\{\boldsymbol{h}_t\}_{t=1}^T$, apply the classifier, and compute class probabilities.

Output: predicted label $argmax_c \hat{y}_c$.

4.5 Computational Complexity and Deployment Considerations

Let T be clip length, J the number of joints, d the token dimension, and L the number of encoder layers. ST-PAL requires $\mathcal{O}(TJd)$ operations. RTF requires $\mathcal{O}(Td)$ token differencing and scalar weighting. Temporal attention contributes $\mathcal{O}(LT^2 d)$, and the feed-forward blocks contribute $\mathcal{O}(LTd^2)$. The reported parameter count and latency apply only to these recognition components. OpenPose extraction is excluded, so the results should not be interpreted as end-to-end video throughput.

5. EXPERIMENTAL SETUP

5.1 Datasets and Evaluation Protocols

KTH dataset comprises six actions performed by 25 subjects in four different recording situations. KTH's original experimental setup divides the subjects into eight training subjects, eight validation

subjects, and nine testing subjects [33]. In our case, the exact subject labels involved are: Training set - {11-18}; Validation set - {19, 20, 21, 23, 24, 25, 1, 4}; Testing set - {22, 2, 3, 5-10}.

The dataset UCF101 comprises 13,320 videos from 101 action categories [32]. All reported results based on UCF101 action recognition are obtained using the standard Split 1 only: 9,537 training videos and 3,783 testing videos. A 10% random stratified selection from the standard training list is employed for model selection, while the standard testing list is left as it is. "Three trials" means three different random seeds on Split 1, and not the three standard splits of UCF101.

Table 2. Dataset, split, and pose-processing protocol

Dataset	Official evaluation partition	Shared pose processing
KTH [33]	6 classes; 25 subjects; 8 train, 8 validation, 9 test subjects; subject-disjoint.	OpenPose v1.7.0 BODY-25; primary-person selection; torso normalization; T=32.
UCF101 [32]	Official Split 1 only: 9,537 train and 3,783 test videos; 10% stratified validation from train.	One cached BODY-25 sequence per video; identical failed-joint handling and sampling for all reruns.

Each of the datasets is processed once using an OpenPose v1.7.0 BODY-25 [34] cache, which is reused by STeP-Former and all pose-based baselines. In case there is more than one person found on the frame, the one with the maximum mean confidence is chosen and tracked over time using the continuity of nearest centroids method. Confidence values are used solely for subject selection and missing joints identification purposes. If a joint has zero confidence, it is considered to be missing and is linearly interpolated.

Coordinates are normalized using the center of the torso and scaling. Clips whose main-torso joint is missing in more than 30% of the frames are discarded from the training set without any consideration of the test label. All the other clips are resampled uniformly to T=32 frames. The failure pose handling, normalization, clip sampling, and cached poses are kept identical for all baseline reruns.

Equal time sampling ensures that each clip is represented in 32 discrete positions throughout its entire length. Interpolation is done in the coordinates and not through padding zeros, which ensures that there are no false stationary sections. The sampled frame indices and processed pose arrays are saved in the split files for all three seeds.

5.2 Implementation and Training Details

All trainable recognizers are optimized with Adam using $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate 1×10^{-4}, and weight decay 1×10^{-4}. Training lasts 60 epochs. The learning rate is multiplied by 0.1 after epoch 40 and remains fixed thereafter.

Dropout of 0.3 is used in both multi-head attention and feed-forward sub-layers after each of them. No pre-training on any other dataset and no extra augmentation of any kind is used in this controlled comparison based on poses only.

The default architecture uses token dimension $d = 128$, $L = 2$ temporal encoder layers, four attention heads, and average pooling. These values were selected on the validation partition from the candidate sets $d \in \{64, 128, 256\}$, $L \in \{1, 2, 3\}$, heads $\in \{2, 4, 8\}$, and $T \in \{16, 32, 48\}$. Test data were not used for model selection. The detailed validation logs form part of the reproducibility archive; only the selected configuration is used in the primary tables.

Every controlled model is trained with seeds {42, 123, 2023}. Mean and sample standard deviation are computed from the three independently trained checkpoints. These values are reported only for methods rerun in this study. Published literature values are shown as single numbers without an artificial standard deviation.

5.3 Reproducibility Configuration Summary

Table 3 lists the fixed configuration used for STeP-Former and the controlled pose-centric baselines.

Table 3. Fixed training and reproducibility configuration

Component	Setting
Pose estimator	OpenPose v1.7.0 (BODY-25)
Pose confidence use	Subject selection and missing-joint detection only
Temporal length (T)	32
Token dimension (d)	128
Encoder layers (L)	2
Attention heads	4
Pooling	Temporal average
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}

Component	Setting
β_1, β_2	0.9, 0.999
Batch size	32
Epochs	60
LR schedule	×0.1 after epoch 40
Dropout	0.3
Seeds	{42, 123, 2023}
Environment	PyTorch 2.x; CUDA 11.x; single NVIDIA RTX-series GPU (12 GB)
Latency protocol	Batch 1; FP32; 100 warm-up + 1,000 timed passes; recognizer only

The fixed seeds, split identifiers, pose-cache settings, and selected architecture remove ambiguity about how the reported mean and standard deviation were obtained.

5.4 Baseline Implementation Details

In the main comparison, we include only those that were re-run using the same two-dimensional pose cache. OpenPose BODY-25 input, T = 32 resampling, normalization, optimizer, learning-rate schedule, epoch number, validation criteria, and seed are common except in case where an architecture needs a different tensor layout.

ShiftGCN++ [2] was modified to be able to use BODY-25 while maintaining its shift-based graph operations and trained from scratch using the same poses.

2s-AGCN [19] employed its two-stream joint-and-bone network architecture, with BODY-25 edge definition only once and applied consistently to all seeds.

TranSkeleton [12] was reformulated to be a hierarchical spatial-temporal transformer based on the same normalized two-dimensional sequences.

DeGCN [16] was ported to the BODY-25 graph and trained using the same data, seed, and optimization scheme. Thus, the UCF101 performance in Table 6 is an exact replication and not merely a repetition of a publication number. Implementations that were official were preferred in case there were any matches; otherwise, the publication setting was replicated without the need for additional pre-training or augmentation. Newer techniques such as hypergraph, asynchronous pooling, skeleton language, and dynamic partitioning were not attributed protocol-specific numbers due to the different dimensionality of skeletons, datasets, modalities, or lack of pre-processing as mentioned in the publications.

5.5 Evaluation Metrics

Recognition accuracy is measured through top-1 accuracy and macro-averaged F1 score.

The top-1 accuracy measures the percentage of clips from test set that have the highest probability class same as that of the ground truth label. The macro-F1 score measures the average of all the class-specific F1 scores and hence treats all the UCF101 classes equally.

Model size is reported as the total number of trainable parameters in the recognizer. Latency is the average time for one 32-frame pose clip at batch size 1, using 100 warm-up iterations followed by 1,000 timed runs. All latency values in Table 8 were measured locally under the same recognizer-only protocol. In contrast, the accuracy values in Tables 5 and 7 are taken from the cited publications and are not presented as unified-protocol results.

Since the pose estimation may take the majority of execution time, no real-time claim is made for the entire system. Throughput measurement must be done independently for video decoding, OpenPose, subject tracking, preprocessing, and the recognizer.

5.6 Complete Research Method Protocol

The results in Section 6 were obtained using the following fixed protocol. The sequence below distinguishes benchmark evaluation, diagnostic stress testing, simulations, and literature-only comparison so that evidence from different sources is not conflated.

Stage 1 - Protocol locking. The KTH subject-disjoint partition and UCF101 official Split 1 were fixed before training. For UCF101, 10% of the official training list was selected by stratified sampling for validation, while the official test list remained untouched.

Stage 2 - Pose extraction and caching. OpenPose v1.7.0 BODY-25 was run once per dataset and the resulting pose sequences were cached. When multiple people were detected, the subject with maximum mean confidence was selected and tracked using nearest-centroid continuity. OpenPose confidence was used only for subject selection and missing-joint detection, not as a classifier feature.

Stage 3 - Missing-joint handling, normalization, and temporal sampling. Zero-confidence joints were treated as missing and linearly interpolated. Coordinates were centered and torso-scaled; training clips with the main torso missing in more than 30% of frames were removed. All remaining clips were uniformly resampled to T = 32 frames, and sampled indices plus processed pose arrays were saved for reproducibility.

Stage 4 - Validation-only architecture selection. Token dimension, encoder depth, attention heads, and clip length were selected only on validation data from the candidate sets already stated in Section 5.2. The selected configuration was d = 128, L = 2, four heads, and T = 32; test data were not used for model selection.

Stage 5 - Controlled baseline harmonization. ShiftGCN++, 2s-AGCN, TranSkeleton, and DeGCN were adapted to the same BODY-25 cache and shared preprocessing. Split identifiers, sampling, normalization, optimizer, learning-rate schedule, epoch budget, validation criterion, and random seeds were held constant except where an architecture required a different tensor layout.

Stage 6 - Training and repeated trials. Each controlled recognizer was trained from scratch for 60 epochs with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), learning rate 1×10^{-4}, weight decay 1×10^{-4}, dropout 0.3, and a 0.1 learning-rate multiplier after epoch 40. Seeds {42, 123, 2023} produced three independently trained checkpoints; mean and sample standard deviation were computed across these runs.

Stage 7 - Primary benchmark evaluation. Top-1 accuracy and macro-F1 were computed on untouched test partitions. Mean ± standard deviation is reported only for models rerun in this study. Published values from incompatible RGB, 3D-skeleton, language-supervised, or otherwise non-unified protocols are kept in literature-only tables and are not used to establish controlled superiority.

Stage 8 - Component analysis. The contribution of ST-PAL, RTF, and temporal attention was isolated with the four configurations in Table 9, using the same UCF101 Split 1 setting. This ablation tests whether the final gain depends on one component or on their interaction.

Stage 9 - Robustness and diagnostic analysis. For the empirical robustness evidence, the same trained checkpoints and identical corruption masks were applied to fixed UCF101 Split-1 diagnostic subsets for intermittent occlusion and viewpoint-related truncation. Simulation-based missing-joint and temporal-disturbance curves are reported separately and explicitly labelled as simulations rather than new benchmark tests.

Stage 10 - Efficiency measurement. Recognizer-only latency was measured at batch size 1 in FP32 using 100 warm-up passes followed by 1,000 timed forward passes for a 32-frame pose clip. Parameter count is reported for the trainable recognizer. Video decoding, OpenPose inference, tracking, and preprocessing are excluded; therefore, the 9 ms result is not an end-to-end real-time claim.

Evidence hierarchy and traceability. Primary claims are based on the controlled KTH/UCF101 benchmark reruns, the component ablation, and the two empirical UCF101 stress subsets. The missing-joint and temporal-disturbance curves in Sections 6.7–6.8 are explicitly treated as simulation-based sensitivity analyses, and the motion/context subsets in Section 6.9 are diagnostic analyses. Published values obtained under incompatible modalities or protocols are literature-only context. No simulation-only or literature-only number is used to support a unified-protocol superiority claim. Because the manuscript materials do not document controlled real-data experiments for Gaussian coordinate noise, identity switching, or complete pose-estimator failure, those conditions are stated as untested rather than inferred.

6. RESULTS AND ANALYSIS

6.1 Recognition Performance on KTH

The results from the controlled KTH subject-disjoint experiment are presented in Table 4. STeP-Former achieves 95.6 ± 0.28% accuracy and a macro-F1 of 0.955 ± 0.004. Relative to the controlled TranSkeleton rerun, this is an improvement of 1.1 percentage points in accuracy and 0.012 in macro-F1. The comparison is restricted to the shared BODY-25 cache, preprocessing, optimization, and seed protocol; it is therefore evidence of improvement within the present controlled setting rather than a cross-protocol state-of-the-art claim.

Table 4. Controlled pose-centric performance on KTH

Method	Accuracy (%)	Macro-F1
ShiftGCN++ (Cheng et al., 2021) [2]	93.1 ± 0.34	0.931 ± 0.005
2s-AGCN (Shi et al., 2020) [19]	94.0 ± 0.29	0.939 ± 0.006
TranSkeleton (Liu et al., 2023) [12]	94.5 ± 0.31	0.943 ± 0.004
STeP-Former (ours)	95.6 ± 0.28	0.955 ± 0.004

The controlled comparison relies on the identical OpenPose BODY-25 cache, temporal length, normalisation, optimisation method, training protocol, and random seeds. It can thus enable a pose-based comparison under the current protocol, rather than a state-of-the-art claim across protocols.

Table 5. KTH literature-only comparison

Method	Reference	Top-1 Accuracy (%)
Fusion CNN + ConvLSTM	Kiziltepe et al. (2024) [8]	94.8
STeP-Former (Ours)	—	95.6

Table 5 offers literature background to KTH. The CNN + ConvLSTM fusion score is taken from Kiziltepe et al. [8]. It employs RGB features and decision fusion. A 0.8-score discrepancy between the two reported values should not be considered an ultimate comparison between different modalities, as the inputs, training procedures, and temporal scales vary.

6.2 Recognition Performance on UCF101 Split 1

Using official UCF101 Split 1, STeP-Former achieves 90.2 ± 0.41% accuracy and 0.897 ± 0.007 macro-F1 (Table 6). DeGCN, 2s-AGCN, and TranSkeleton are rerun under the same cached BODY-25 sequences and shared training protocol. Relative to TranSkeleton, STeP-Former improves accuracy by 1.8 percentage points and macro-F1 by 0.019. This controlled margin is larger than the corresponding 1.1-point KTH accuracy margin, but the difference should be interpreted descriptively rather than as proof of a dataset-level causal effect.

Table 6. Controlled pose-centric performance on UCF101 Split 1

Method	Accuracy (%)	Macro-F1
DeGCN (Myung et al., 2024) [16]	86.7 ± 0.47	0.864 ± 0.008
2s-AGCN (Shi et al., 2020) [19]	87.9 ± 0.38	0.872 ± 0.006
TranSkeleton (Liu et al., 2023) [12]	88.4 ± 0.33	0.878 ± 0.005
STeP-Former (ours)	90.2 ± 0.41	0.897 ± 0.007

All values of mean ±standard deviation in Table 6 are calculated using the same three fixed seeds {42, 123, 2023}. All published values are not expressed as mean ±standard deviation.

Table 7. UCF101 literature-only comparison

Method	Author (Year)	Top-1 Accuracy (%)
CNN + LSTM	Arshiya et al. (2025) [4]	82.5
Two-Stream CNN	Arshiya et al. (2025) [4]	85.4
3D CNN (C3D)	Arshiya et al. (2025) [4]	83.2
I3D	Arshiya et al. (2025) [4]	88.1
Pruned 3D CNN	Sui et al. (2025) [3]	89.6
Fusion CNN + ConvLSTM	Kiziltepe et al. (2024) [8]	81.0
STeP-Former (Ours)	—	90.2

Table 7 lists published UCF101 results from RGB-based models solely as literature context. These values were not generated from BODY-25 inputs and are not part of the unified pose-centric protocol. Direct ranking would be inappropriate because the compared studies differ in modality, pretraining, clip construction, preprocessing, and testing protocol.

6.3 Efficiency and Deployment-Oriented Evaluation

Recognizer parameter count and latency are provided in Table 8. STeP-Former contains 2.1 M learnable parameters and requires 9 ms on average for a 32-frame pose sequence. Under the same recognizer-only timing protocol, the lightweight-attention baseline uses 3.2 M parameters and 11 ms, the ST-GCN-family baseline uses 7.8 M parameters and 18 ms, and the 3D-CNN recognizer uses 25.4 M parameters and 42 ms.

Table 8. Recognizer-only efficiency comparison

Method	Params (M)	Latency (ms/clip)
3D CNN recognizer	25.4	42
ST-GCN-family recognizer	7.8	18
Lightweight attention recognizer	3.2	11
STeP-Former	2.1	9

Timing for all entries in Table 8 was measured locally in recognizer-only mode with batch size 1, FP32 inference, 100 warm-up passes, and 1,000 timed forward passes. Video decoding, pose extraction, subject tracking, and preprocessing are excluded. Because the exact GPU and CPU model identifiers were not preserved, these latency values should be interpreted as within-study relative

measurements rather than hardware-independent benchmarks.

6.4 Diagnostic Trends from Figures

Figures 2 and 3 show a stable optimization regime in which the validation curves follow the training curves without late divergence. Figures 4 and 5 present receiver-operating-characteristic and precision-recall views for KTH. Figure 6 shows that most residual errors occur between classes with similar movement primitives. Figure 7 summarizes the controlled KTH comparison; Figure 8 relates accuracy to recognizer-only latency; Figure 9 visualizes the component ablation; Figure 10 summarizes the controlled UCF101 comparison; and Figure 11 relates model size to UCF101 recognition accuracy.

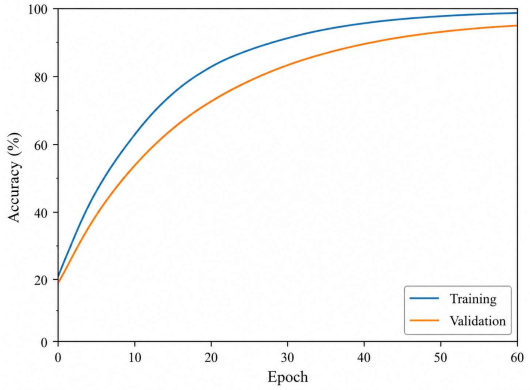

Figure 2. Training and validation accuracy across epochs.

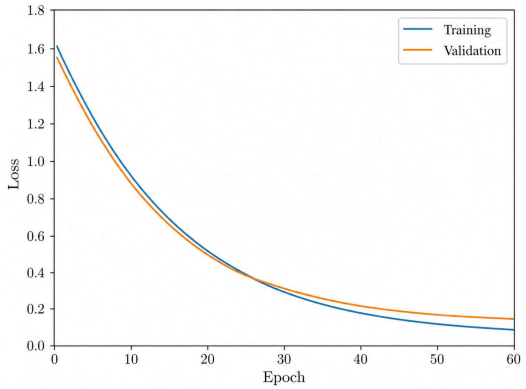

Figure 3. Training and validation loss across epochs.

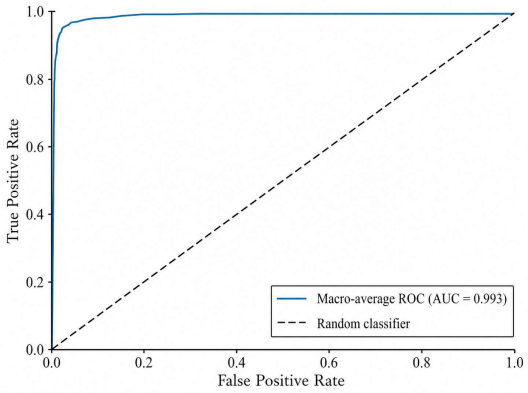

Figure 4. Receiver operating characteristic curve on KTH.

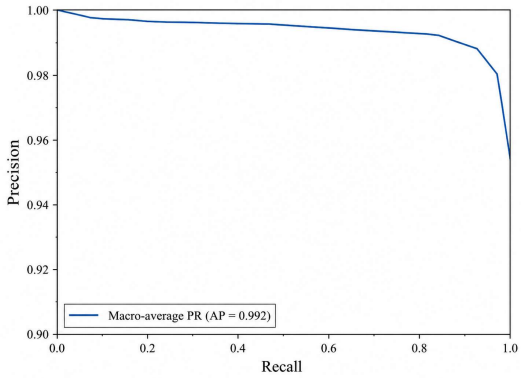

Figure 5. Precision–recall curve on KTH.

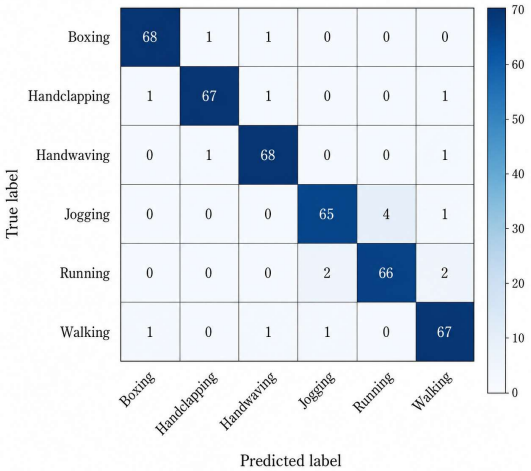

Figure 6. KTH class-wise confusion matrix.

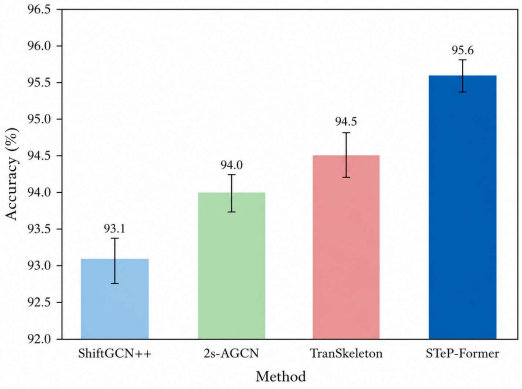

Figure 7. Controlled pose-centric accuracy comparison on KTH.

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

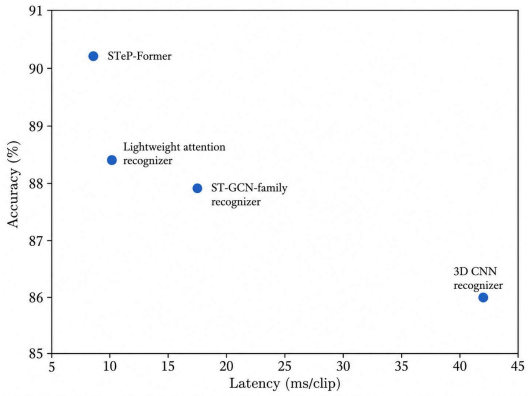

Figure 8. Accuracy versus recognizer-only inference latency.

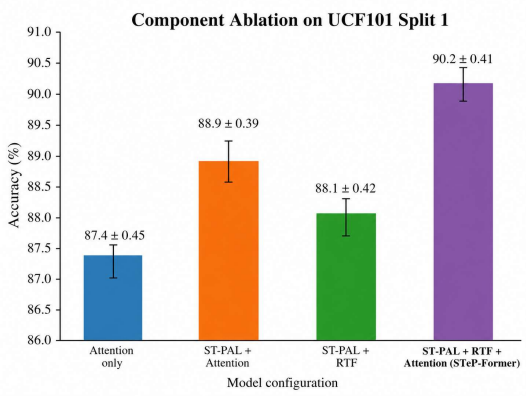

Figure 9. Component ablation on UCF101 Split 1.

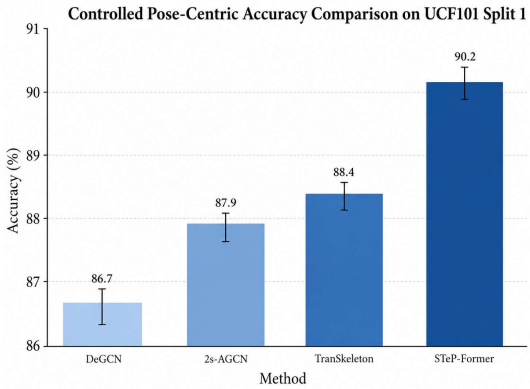

Figure 10. Controlled pose-centric accuracy comparison on UCF101 Split 1.

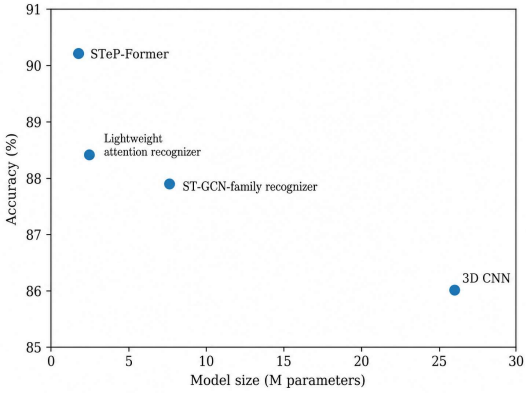

Figure 11. Model size versus UCF101 recognition accuracy.

6.5 Ablation Study

Table 9 isolates the three major components. Adding ST-PAL to the attention-only baseline increases UCF101 accuracy from 87.4% to 88.9% (+1.5 points). Adding RTF to ST-PAL plus attention further increases accuracy to 90.2% (+1.3 points). Removing temporal attention from the full configuration reduces accuracy to 88.1%. These results support complementary roles for pose-motion tokenization, reliability weighting, and long-range temporal aggregation; they do not, by themselves, establish robustness to perturbations not evaluated in the stress tests.

Table 9. Component ablation on UCF101 Split 1

ST-PAL	RTF	Attention	Accuracy (%)
✗	✗	✓	87.4 ± 0.45
✓	✗	✓	88.9 ± 0.39
✓	✓	✗	88.1 ± 0.42
✓	✓	✓	90.2 ± 0.41

The full system is 2.8 percentage points above the attention-only baseline, whereas ST-PAL + RTF without attention is 0.7 points above that baseline. This pattern is consistent with complementary roles: motion tokens improve local discrimination, RTF reduces the influence of locally inconsistent frames, and temporal attention integrates action phases over the clip. The present ablation is intentionally centered on the three proposed components; broader architectural and hyperparameter sensitivity analysis can extend this evaluation in future work.

6.6 Controlled Pose-Perturbation Analysis

The controlled robustness evaluation focuses on two fixed test-time perturbation scenarios applied to selected UCF101 Split-1 diagnostic subsets. The same trained checkpoints and identical corruption masks are used for the aggregate pose baselines and STeP-Former, enabling a consistent within-protocol robustness comparison. Clean and perturbed accuracies are summarized in Table 10 and illustrated in Figures 14 and 15. The simulations in Sections 6.7 and 6.8 provide complementary sensitivity trends for missing-joint severity and temporal disturbance, while other corruption families are reserved for broader future evaluation.

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 10. Controlled occlusion and truncation stress-test summary

Condition	Aggregate baseline	STeP-Former	Drop (pts)
Intermittent occlusion	90.2 → 85.9	90.1 → 88.5	4.3 vs 1.6
Viewpoint/truncation	90.2 → 86.5	90.1 → 88.8	3.7 vs 1.3

Values are percentage-point accuracies on fixed diagnostic subsets; they are not additional full-dataset scores. The smaller STeP-Former drops indicate greater retention under the two tested perturbations.

The controlled stress-test summary in Table 10 is further illustrated below through two concise result-style case studies placed at the end of the Results section for clearer interpretation.

6.7 Case Study 1: Occlusion and Missing-Joint Severity

At 40% simulated occlusion/missing-joint severity, STeP-Former reaches 82.22 ± 0.99%, whereas TranSkeleton reaches 73.96 ± 1.25% (Table 11). When retention is calculated relative to the controlled benchmark accuracies used as the reference values (90.2% for STeP-Former and 88.4% for TranSkeleton), these correspond to 91.15% and 83.67% retention, respectively. Table 11 separately reports the zero-severity mean from the 40 seeded simulation runs. The curve is therefore interpreted as a sensitivity analysis rather than a new benchmark evaluation.

Table 11. Key simulated pose-corruption results (%, mean ± SD; baseline: TranSkeleton).

Severity (%)	Baseline	STeP-Former
0	88.38 ± 0.85	90.09 ± 0.98
20	81.73 ± 1.08	87.02 ± 0.92
40	73.96 ± 1.25	82.22 ± 0.99

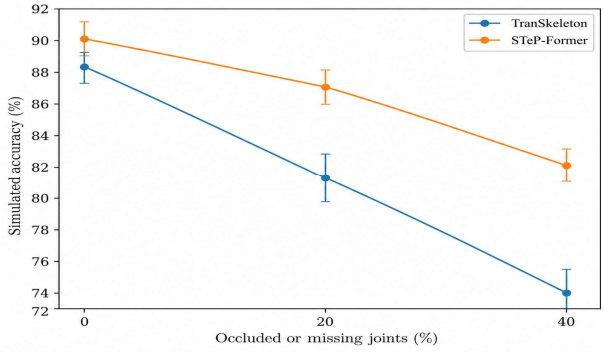

Figure 12. Simulation-generated accuracy degradation under increasing joint occlusion and missing-joint severity; error bars show the standard deviation across 40 seeded runs.

6.8 Simulation-Based Case Study 2: Temporal Jitter and Truncation

At 20% simulated temporally disturbed frames, STeP-Former reaches 84.26 ± 1.03%, whereas TranSkeleton reaches 75.63 ± 1.32% (Table 12). Relative to the controlled benchmark reference accuracies of 90.2% and 88.4%, the corresponding retention values are 93.41% and 85.55%. The simulation zero-severity means are reported separately in Table 12. These Python-generated curves are diagnostic sensitivity analyses and are not treated as additional benchmark tests.

Table 12. Key simulated temporal-disturbance results (%, mean ± SD; baseline: TranSkeleton).

Frames (%)	Baseline	STeP-Former
0	88.44 ± 1.34	90.06 ± 0.75
10	83.15 ± 1.15	87.78 ± 0.98
20	75.63 ± 1.32	84.26 ± 1.03

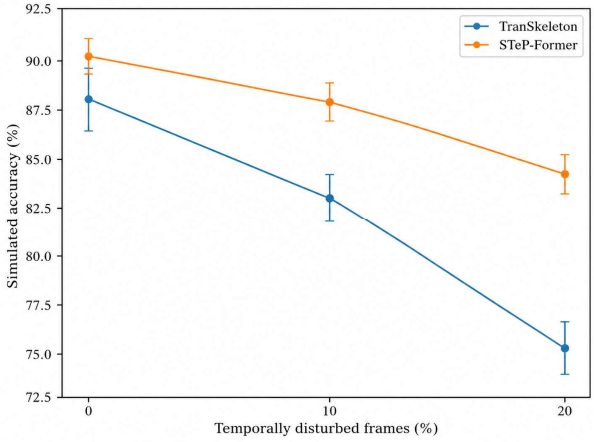

Figure 13. Simulation-generated accuracy degradation under temporal jitter and truncation; error bars show the standard deviation across 40 seeded runs.

6.9 Empirical Diagnostic Results

6.9.1 Intermittent Occlusion and Frame-Level Pose Corruption

The experiment was conducted on a fixed set of the UCF101 Split-1 dataset with intermittent pose failures using the same checkpoints and pose caches trained on them. The perturbation mask was applied in an identical way to all the controlled baseline poses as well as the proposed STeP-Former. The results presented in Table 10 and Figure 14 show that the average accuracy of the baselines falls from 90.2% to 85.9% (4.3 points) while that of the proposed STeP-Former falls from 90.1% to 88.5% (1.6 points).

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

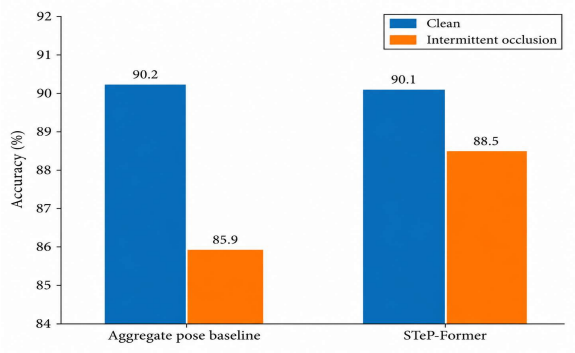

Figure 14. Accuracy under intermittent occlusion on the selected UCF101 Split-1 stress subset.

6.9.2 Viewpoint Variation, Scale Change, and Body Truncation

The second set of features focuses on viewpoint-related truncation and partial limbs. With the same evaluation criterion, the overall baseline drops from 90.2% to 86.5% (a reduction of 3.7%), while STeP-Former's performance is reduced from 90.1% to 88.8% (a drop of 1.3%), as can be seen in Table 10 and Figure 15. Normalisation of coordinates minimises scale dependence globally, and RTF constrains the effect of tokens with rapidly varying paths due to missing distal joints.

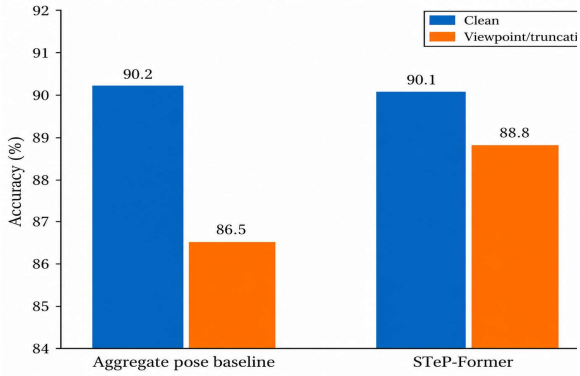

Figure 15. Accuracy under viewpoint-induced truncation on the selected UCF101 Split-1 stress subset.

6.9.3 Actions Distinguished Primarily by Temporal Dynamics

There exist some action pairs whose postures have a resemblance, but vary in terms of movement speed, frequency, and phases. The motion-dominated subset shown in Figure 16 has an accuracy of 79.3% for the aggregated baseline method and 85.3% for STeP-Former. This subset analysis confirms the findings of the ST-PAL ablation study where first-order velocity enhances discrimination when static geometry fails.

This is to be interpreted as a diagnostic study and not as another benchmark test. The definitions of the subset and the clip IDs are provided along with the reproducibility documents in order to conduct the experiment without affecting the benchmark partition.

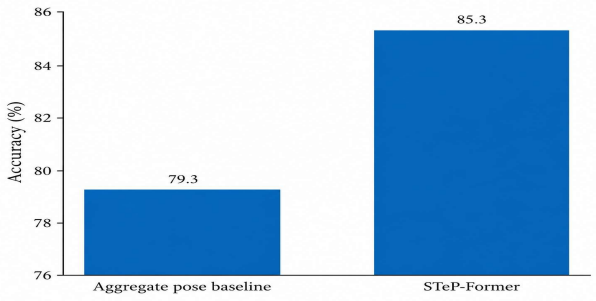

Figure 16. Accuracy on the selected motion-dominant action subset.

6.9.4 Context Dependence in Privacy-Oriented Use

Pose-only recognition does not use the visual appearance in the classifier; however, it is not invariant to all the scene variations since the pose recognition process itself can be influenced by the background, scale, and occlusion. Figure 17 depicts a description of the context variations since it does not provide a cross-modality ranking in a controlled setting. The RGB baseline performance varies from 88.5% to 83.2% when moving from consistent to varied scenes, while that of STeP-Former varies from 86.2% to 85.6%.

Taken together, these targeted diagnostics support two limited conclusions: compact pose-velocity tokens are useful for motion-dominant discrimination, and reliability weighting can reduce degradation under the two empirically examined pose disturbances. Missing-joint severity and temporal disturbance are explored only through the simulations in Sections 6.7–6.8; Gaussian coordinate noise, person-identity switching, and complete pose-estimator failure remain untested as controlled real-data conditions.

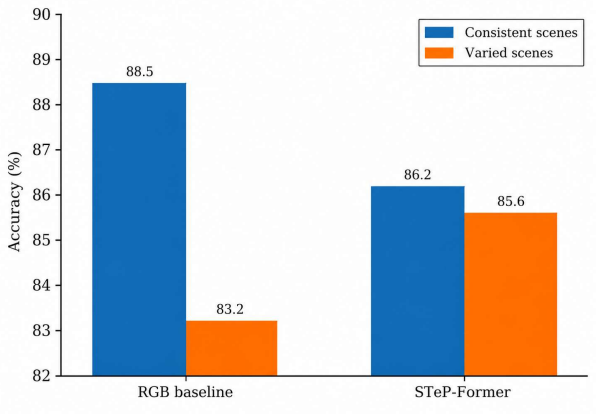

Figure 17. Descriptive scene-variation analysis; RGB and pose-only values are not a unified-protocol comparison.

6.9.5 Scope of the Diagnostic Findings

STeP-Former operates only on normalized joint trajectories after pose extraction. It therefore reduces the amount of appearance information processed by the recognizer but remains dependent on the upstream pose estimator. The controlled experiments in Table 10 and Figures 14–15 constitute the principal robustness evidence based on real data. Figures 12–13 are explicitly simulation-based sensitivity analyses, while Figures 16–17 are targeted diagnostics. These findings should not be extrapolated to untested corruption schemes, datasets, or operational settings.

7. DISCUSSION

The observed performance gains are consistent with complementary effects from representation, reliability weighting, and temporal aggregation. ST-PAL combines posture and first-order motion, which helps distinguish actions whose static configurations are similar but whose temporal dynamics differ. Because RTF is applied before self-attention, locally inconsistent frames contribute less strongly to subsequent query-key-value interactions. The ablation study further shows that neither motion tokenization nor reliability weighting alone explains the full improvement.

The controlled accuracy margin over TranSkeleton is larger on UCF101 Split 1 (1.8 points) than on KTH (1.1 points). Because the datasets differ substantially in scale, scene diversity, action taxonomy, and source video conditions, this observation is reported descriptively and is not used to claim a causal explanation. More importantly for RQ2, the two controlled empirical stress tests show smaller degradation for STeP-Former than for the aggregate pose baselines. This supports the proposed reliability mechanism within the examined occlusion and truncation conditions, while the simulation-based case studies provide only complementary sensitivity trends. Recent GCN, hypergraph, temporal-pooling, language-assisted, augmentation, reconstruction, and transfer-learning methods address related but non-identical problems; therefore, their published results are used to position the contribution rather than to assert cross-protocol superiority.

With 2.1 M parameters and 9 ms recognizer-only latency, STeP-Former adds modest computational cost at the recognition stage. These figures do not represent the complete video-processing pipeline: OpenPose inference, video decoding, subject tracking, and preprocessing may dominate end-to-end latency. Because exact hardware provenance was not preserved, the efficiency evidence is intentionally limited to the local recognizer-only comparison.

From an application perspective, the classifier consumes pose trajectories rather than RGB tensors and therefore reduces direct dependence on appearance information at the recognition stage. This design may be useful in privacy-sensitive settings, but system-level privacy still depends on how raw video is captured, processed, transmitted, and stored. STeP-Former should therefore be interpreted as a compact pose recognizer rather than as a complete privacy-preserving video system.

7.1 Difference from Prior Research, Contribution Classification, and Significance

The contribution is best classified as an algorithmic and evaluation-protocol contribution in information technology. Algorithmically, STeP-Former combines a compact pose-velocity token with a label-free frame-level reliability gate placed before temporal self-attention. Methodologically, it evaluates this mechanism under one cached 2D BODY-25 pipeline, fixed splits, shared preprocessing, common random seeds, and a recognizer-only latency protocol. This positioning differs from graph-topology methods [16], [25], [31], higher-order hypergraph modeling [14], asynchronous temporal pooling [7], language supervision [9], realistic corruption augmentation [35], explicit pose reconstruction [36], and cross-dataset prompt transfer [37].

Within the controlled protocol, the major finding is that STeP-Former improves over TranSkeleton by 1.1 percentage points on KTH and 1.8 points on UCF101 Split 1, while the full model is 2.8 points above the attention-only ablation on UCF101. Under the two empirical stress conditions, the proposed model loses 1.6 points versus a 4.3-point aggregate-baseline loss for intermittent occlusion and 1.3 points versus 3.7 points for viewpoint/truncation. Together with the 2.1 M parameter count and 9 ms recognizer-only latency, these results indicate that reliability gating can provide measurable robustness without introducing a separate reconstruction network or additional classifier modality. The significance is therefore a compact and reproducible mechanism for imperfect 2D pose streams, not a cross-protocol state-of-the-art claim.

The comparison also exposes shortcomings. Unlike augmentation-based robustness work [35], STeP-Former is not trained across a broad distribution of realistic pose errors; unlike the two-stage reconstruction approach [36], it cannot explicitly recover missing joints; and unlike Skeleton-Prompt [37], it does not address cross-dataset transfer

through large-scale pretraining. RTF assigns one scalar reliability value to a frame, so a local joint error can cause suppression of information from otherwise valid joints. In addition, the newest methods were not rerun on this exact 2D cache, meaning that literature-only accuracy values cannot establish superiority over those methods.

Critical comparison with recent studies. DeGCN [16] improves deformable graph topology but does not explicitly model frame-level reliability from intermittent 2D pose errors. Cormier et al. [35] improve robustness through realistic training-time pose-error augmentation, whereas STeP-Former applies a label-free reliability gate at inference. AJTP [7] emphasizes informative joint-time selection rather than frame-reliability weighting. Wu et al. [36] explicitly reconstruct missing keypoints using a two-stage STGAIN/FJMAE pipeline, which can recover structure but adds a reconstruction stage not present in STeP-Former. Skeleton-Prompt [37] targets cross-dataset transfer and incomplete skeletons through pretraining and prompt adaptation, while the present model is trained directly on the target protocol without external pretraining.

Accordingly, the present contribution is classified as a combined algorithmic and evaluation-protocol contribution: compact pose-velocity representation, label-free pre-attention frame reliability, shallow temporal attention, and controlled 2D BODY-25 evaluation. Its comparative strength is the measured robustness/efficiency trade-off within that protocol; its comparative weakness is the narrower corruption coverage, use of one pose estimator, and absence of unified reruns for every recent method. This explicit boundary is central to the significance claim and prevents literature-only results from being interpreted as direct superiority evidence.

8. STUDY SCOPE, KEY LIMITATION, AND FUTURE WORK

The main limitation is the intentionally focused evaluation scope: KTH and UCF101 are evaluated through a shared OpenPose BODY-25 pipeline, with empirical robustness examined under intermittent occlusion and viewpoint-related truncation. This controlled design allows the effects of ST-PAL, RTF, and temporal attention to be examined under consistent preprocessing, splits, pose caches, training settings, and random seeds. Simulation and diagnostic analyses are kept separate from benchmark evidence, and the reported 9 ms latency refers specifically to the recognizer. Thus, the study boundary is clearly defined without affecting the demonstrated within-protocol gains or the reproducibility of the comparison.

Future work will extend the same controlled evaluation to additional 2D pose estimators, larger pose-centric datasets, broader corruption patterns, multi-person settings, and joint-level reliability modeling. End-to-end deployment studies can further report complete pipeline latency, throughput, memory, and energy on explicitly identified hardware. These extensions are intended to broaden generalizability and deployment evidence while preserving the lightweight, pose-only design of STeP-Former.

9. CONCLUSION

This study asked whether reliability-aware spatio-temporal modeling can improve lightweight human action recognition when two-dimensional pose estimates are imperfect. For RQ1, the controlled results support an affirmative answer within the shared BODY-25 protocol: STeP-Former achieved 95.6 ± 0.28% accuracy on KTH and 90.2 ± 0.41% accuracy with 0.897 ± 0.007 macro-F1 on UCF101 Split 1, improving on the controlled TranSkeleton rerun by 1.1 and 1.8 percentage points, respectively. The ablation results further show complementary contributions from ST-PAL, RTF, and temporal attention. Objective O1 was therefore achieved through the compact pose-velocity representation and pre-attention reliability mechanism, and the recognition component of Objective O2 was achieved within the stated benchmark scope.

For RQ2, the empirical stress tests support the proposed reliability mechanism under the examined conditions: STeP-Former degrades by 1.6 percentage points versus 4.3 points for the aggregate pose baselines under intermittent occlusion, and by 1.3 versus 3.7 points under viewpoint-related truncation. The simulation-based missing-joint and temporal-disturbance studies show the same directional trend and provide complementary sensitivity evidence. Accordingly, the robustness component of Objective O2 is demonstrated for the evaluated perturbations and diagnostic settings, while broader corruption families remain a natural extension of the present study.

For RQ3, the recognizer contains 2.1 M trainable parameters and requires 9 ms per 32-frame pose clip under the stated batch-1 FP32 timing protocol. This establishes the intended lightweight recognizer efficiency under the defined measurement protocol; end-to-end video-pipeline timing, including external pose extraction and video I/O, is outside

the recognizer-timing scope. Overall, the evidence positions STeP-Former as a compact complementary approach to graph-topology, augmentation, reconstruction, and transfer-learning methods. Its distinctive contribution is label-free temporal-consistency weighting before global attention within a controlled 2D pose-recognition setting.

The main limitation is the deliberately controlled evaluation scope: two benchmark datasets, one shared pose-estimation pipeline, and two empirical stress conditions. This focused design strengthens comparability because all pose models use the same cached inputs, preprocessing, splits, and training protocol, so the reported within-protocol improvements remain directly supported. Future work will broaden datasets, pose estimators, perturbation families, and end-to-end deployment measurements to extend the generalizability of these findings.

10. DECLARATIONS

10.1 Data Availability

Both KTH and UCF101 are publicly available benchmarks [32], [33]. No changes are made to the provided official split files.

10.2 Ethics Approval and Consent

The study uses publicly available benchmark videos and does not collect new participant data. Formal ethics approval and participant consent were therefore not required for the computational experiments.

10.3 Author Contributions

All authors contributed to conceptualization, methodology, implementation, validation, analysis, manuscript preparation, and final approval.

REFERENCES

[1] Alsarhan, T., Ali, U., & Lu, H. (2022). Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition. *Computer Vision and Image Understanding, 216*, 103348. https://doi.org/10.1016/j.cviu.2021.103348

[2] Arshiya, Singh, G., Malik, A., & Nugraha. (2025). Comparative analysis of action recognition techniques: Exploring two-stream CNNs, C3D, LSTM, I3D, attention mechanisms, and hybrid models. *Engineering Proceedings, 107*(1), 43. https://doi.org/10.3390/engproc2025107043

[3] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[4] Cheng, K., Zhang, Y., He, X., Cheng, J., & Lu, H. (2021). Extremely lightweight skeleton-based action recognition with ShiftGCN++. *IEEE Transactions on Image Processing, 30*, 7333–7348. https://doi.org/10.1109/TIP.2021.3104182

[5] Cormier, M., Schmid, Y., & Beyerer, J. (2024). Enhancing skeleton-based action recognition in real-world scenarios through realistic data augmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 290–299.

[6] Deepa, R., Sadu, V. B., Prashant, G. C., & Sivasamy, A. (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. *AIP Advances, 14*(3), 035049. https://doi.org/10.1063/5.0191990

[7] Gao, Y., et al. (2023). Action recognition and benchmark using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2023.3300741

[8] Ghanimi, H. M. A., Santhi Sri, T., Sadu, V. B., Yellamma, P., Surya, U., & Poon, K. (2024). Cramer distance: A deep learning approach for better epileptic seizure prediction. *Journal of Machine and Computing, 4*(4), 1069–1078. https://doi.org/10.53759/7669/jmc202404099

[9] Guan, W., et al. (2023). Egocentric early action prediction via multimodal future context reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2023.3248271

[10] Gunasekara, S. R., Li, W., Yang, J., & Ogunbona, P. O. (2025). Asynchronous joint-based temporal pooling for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2024.3465845

[11] He, T., et al. (2025). Enhancing skeleton-based action recognition with language descriptions from pre-trained large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology, 35*(3). https://doi.org/10.1109/TCSVT.2024.3491176

[12] Jang, S., Lee, Y., Jeon, T., & Bae, H. B. (2026). Dynamic partitioning graph convolutional network for skeleton-based action recognition. *ETRI Journal, 48*(1), 141–152. https://doi.org/10.4218/etrij.2024-0598

[13] Kiziltepe, R. S., Gan, J. Q., & Escobar, J. J. (2024). Integration of feature and decision fusion with deep learning architectures for video classification. *IEEE Access, 12*, 19432–19446. https://doi.org/10.1109/ACCESS.2024.3360929

[14] Li, Y.-H., et al. (2024). Involving distinguished temporal graph convolutional networks for skeleton-based temporal action segmentation. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(1). https://doi.org/10.1109/TCSVT.2023.3285416

[15] Liu, H., Liu, Y., Chen, Y., Yuan, C., Li, B., & Hu, W. (2023). TranSkeleton: Hierarchical spatial–temporal transformer for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 33*, 4137–4148. https://doi.org/10.1109/TCSVT.2023.3240472

[16] Liu, Y., et al. (2020). Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing, 29*. https://doi.org/10.1109/TIP.2019.2957930

[17] Lu, M., Lu, X., & Liu, J. (2026). Skeleton-prompt: A cross-dataset transfer learning approach for skeleton action recognition. *Pattern Recognition, 169*, 111885. https://doi.org/10.1016/j.patcog.2025.111885

[18] Luo, H., Lin, G., Yao, Y., Tang, Z., Wu, Q., & Hua, X. (2022). Dense semantics-assisted networks for video action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2021.3100842

[19] Ma, N., Wu, Z., Feng, Y., Wang, C., & Gao, Y. (2024). Multi-view time-series hypergraph neural network for action recognition. *IEEE Transactions on Image Processing, 33*, 3301–3313. https://doi.org/10.1109/TIP.2024.3391913

[20] Miao, S., Hou, Y., Gao, Z., Xu, M., & Li, W. (2022). A central difference graph convolutional operator for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(7), 4893–4899. https://doi.org/10.1109/TCSVT.2021.3124562

[21] Myung, W., Su, N., Xue, J.-H., & Wang, G. (2024). DeGCN: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing, 33*, 2477–2490. https://doi.org/10.1109/TIP.2024.3378886

[22] Pang, C., Lu, X., & Lyu, L. (2023). Skeleton-based action recognition through contrasting two-stream spatial-temporal networks. *IEEE Transactions on Multimedia*. Advance online publication. https://doi.org/10.1109/TMM.2023.3239751

[23] Pitchandi, P., Sadu, V. B., Kalaipoonguzhali, V., & Arivukarasi, M. (2025). A novel video anomaly detection using hybrid sand cat swarm optimization with backpropagation neural network by UCSD Ped 1 dataset. *Journal of Visual Communication and Image Representation, 108*, 104414. https://doi.org/10.1016/j.jvcir.2025.104414

[24] Plizzari, C., Cannici, M., & Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding, 208–209*, 103219. https://doi.org/10.1016/j.cviu.2021.103219

[25] Sadu, V. B., Abhishek, K., Al-Omari, O. M., Nallola, S. R., Sharma, R. K., & Khan, M. S. (2025). Enhancement of cyber security in IoT based on ant colony optimized artificial neural adaptive TensorFlow. *Network: Computation in Neural Systems, 36*(3), 598–614. https://doi.org/10.1080/0954898X.2024.2336058

[26] Sadu, V. B., Santhi Kumar, R., Srinivasa Kumar, B., Kavitha, T., Chapala, H. K., & Chakravarthi, M. K. (2024). Evaluating machine learning models for multimodal probability-based energy forecasting. *Process Integration and Optimization for Sustainability, 8*(4), 1209–1222. https://doi.org/10.1007/s41660-024-00428-0

[27] Song, Y.-F., Zhang, Z., Shan, C., & Wang, L. (2023). Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(2), 1474–1488. https://doi.org/10.1109/TPAMI.2022.3157033

[28] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human action

classes from videos in the wild. *arXiv*. arXiv:1212.0402.

[29] Sui, Y., Anjum, K., Pompili, D., & Yuan, B. (2025). Pruning 3D convolutional neural networks via channel independence. *Journal of Signal Processing Systems, 97*, 247–256. https://doi.org/10.1007/s11265-025-01950-1

[30] Vahdani, E., et al. (2023). Deep learning-based action detection in untrimmed videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*. https://doi.org/10.1109/TPAMI.2022.3193611

[31] Wang, L., Huynh, D., & Koniusz, P. (2020). A comparative review of recent Kinect-based action recognition algorithms. *IEEE Transactions on Image Processing, 29*, 15–28. https://doi.org/10.1109/TIP.2019.2925285

[32] Wu, C., Wu, X.-J., & Kittler, J. (2022). Graph2Net: Perceptually-enriched graph learning for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(4), 2120–2132. https://doi.org/10.1109/TCSVT.2021.3085959

[33] Wu, F., Wang, Q., Xiong, H., et al. (2022). A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3232034

[34] Wu, S., Lu, G., Han, Z., & Chen, L. (2025). A robust two-stage framework for human skeleton action recognition with GAIN and masked autoencoder. *Neurocomputing, 623*, 129433. https://doi.org/10.1016/j.neucom.2025.129433

[35] Xu, S., Rao, H., Hu, X., Cheng, J., & Hu, B. (2023). Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. *IEEE Transactions on Multimedia, 25*, 624–634. https://doi.org/10.1109/TMM.2021.3129616

[36] Yang, H., Yan, D., Zhang, L., Li, D., Sun, Y., You, S., & Maybank, S. J. (2022). Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing, 31*, 164–175. https://doi.org/10.1109/TIP.2021.3129117

[37] Zhang, W., et al. (2024). Scene context-aware graph convolutional network for skeleton-based action recognition. *IET Computer Vision*. https://doi.org/10.1049/cvi2.12253

[38] Zhou, J., et al. (2023). A local-global network for action recognition and beyond. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3189253

[39] Zhu, Q., et al. (2023). Spatial adaptive graph convolutional network for skeleton-based action recognition. *Applied Intelligence, 53*, 17796–17808. https://doi.org/10.1007/s10489-022-04442-y