Journal of Information Systems Engineering and Management
2025, 10(29s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

STSP-Net: A Spatial-Temporal Skeletal Perception Network for Robust 3D Pose Estimation in Children's Sports

Wenyue Liu[1], DENISE KOH CHOON LIAN[1*], Zhihao Zhang[1], Jianguo Qiu[2] and Lili Wang[1]
[1] Faculty of Education, Universiti Kebangsaan Malaysia, Bangi Malaysia
[2] Ludong University, Yantai, Shandong, China
Corresponding author: DENISE KOH CHOON LIAN[1]
Email address: denise.koh@ukm.edu.my

ARTICLE INFO

Received: 30 Dec 2024

Revised: 19 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Introduction: Children's sports motion pose estimation has significant applications in sports training, health monitoring, and rehabilitation assessment. However, existing 3D pose estimation methods still face challenges in sports scenarios, including insufficient stability in keypoint detection, unreasonable 3D structures, and a lack of temporal consistency in motion trajectories. These issues lead to poor robustness in pose prediction under high-speed motion and occlusion conditions.

Objectives: To address the limitations of current 3D pose estimation methods, this paper aims to propose a novel framework that enhances the stability, structural plausibility, and temporal consistency of pose estimation in dynamic and complex children's sports scenarios.

Methods: This paper proposes a novel 3D pose estimation framework, STSP-Net (Spatial-Temporal Skeletal Perception Network), which integrates 2D keypoint detection, skeletal structure modeling, and temporal information modeling. Specifically: The **Efficient Keypoint Detection Module (EPE-Module)** employs a motion-region adaptive enhancement mechanism to improve keypoint detection accuracy and reduce jitter. The **Graph-based Skeletal Representation Module (GSR-Module)** constructs a human skeleton graph and utilizes a graph attention mechanism to optimize spatial relationships and ensure physical plausibility. The **Temporal Motion Perception Module (TMP-Module)** adopts a cross-attention mechanism to capture long-term motion trends and applies global temporal constraints to enhance smoothness and consistency.

Results: Experimental results demonstrate that STSP-Net achieves the lowest MPJPE of 48.5 mm on Human3.6M and 49.6 mm on ChildPlay, reducing error by 2.6% and 3.1% compared to the best baseline. It also achieves the lowest TS values of 3.3 mm/s^2 and 3.4 mm/s^2, ensuring smoother motion trajectories. Furthermore, STSP-Net maintains stable pose estimation in high-speed motion and occlusion scenarios, consistently outperforming existing methods.

Conclusions: STSP-Net effectively addresses the core challenges in children's sports motion pose estimation by improving keypoint detection stability, enforcing 3D skeletal consistency, and enhancing temporal smoothness. It offers a robust solution for practical applications in sports, health, and rehabilitation domains.

Keywords: Children’s Pose Estimation, Sports Motion Analysis, Computer Vision.

INTRODUCTION

Children's sports motion pose estimation is an important interdisciplinary research area in computer vision and sports science, with broad applications in sports training, health monitoring, and rehabilitation assessment[1]. Accurate pose estimation enables coaches and teachers to analyze children's dynamic postures in sports in real-time, providing scientific training guidance and effectively preventing injuries caused by poor movement habits. However, traditional pose estimation methods often rely on annotated data or wearable sensors, which not only increase the cost of data acquisition but also limit their applicability in natural environments. In recent years, deep learning has made significant advancements in 2D/3D human pose estimation, making real-time pose estimation based on monocular video feasible[2]. Nevertheless, due to the diversity of children's motion postures, existing methods still face numerous challenges in sports scenarios, including significant pose variations across different movement patterns, difficulty in accurately reconstructing 3D keypoints with monocular cameras, and occlusions affecting certain joints. Therefore, developing an efficient and robust children's sports motion pose estimation method can provide crucial technological support for sports monitoring, physical education, and youth health management[3].

Although deep learning techniques have made considerable progress in pose estimation in recent years, existing methods still have limitations. Directly regressing 3D keypoints from monocular images often struggles to adapt to complex motion scenarios, is highly susceptible to viewpoint variations and occlusions, and suffers from weak generalization ability[4]. Methods that infer 3D poses from 2D keypoints offer improved stability, but they heavily rely on the accuracy of keypoint detection and lack explicit modeling of temporal consistency in motion trajectories, leading to large prediction errors in fast-motion scenarios. Furthermore, most existing approaches are based on Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks, which struggle to fully capture global information, thereby limiting their motion trajectory prediction accuracy[5][6]. At the same time, human poses inherently exhibit a graph-structured topology, yet current methods rarely explicitly model the relationships between joints, restricting the model's ability to assess pose plausibility. Therefore, how to achieve real-time 3D keypoint estimation while improving motion trajectory temporal consistency and enhancing skeletal topology modeling remains an urgent problem to be addressed.

To overcome the aforementioned limitations, we propose a novel framework—STSP-Net (Spatial-Temporal Skeletal Perception Network)—which integrates temporal modeling and skeletal structure modeling to achieve high-precision and stable 3D pose estimation. STSP-Net consists of three core modules: Efficient Pose Extraction (EPE-Module), Graph-Based Skeletal Representation (GSR-Module), and Temporal Motion Perception (TMP-Module)[7][8]. These modules are designed to enhance keypoint detection stability, improve the physical consistency of 3D poses, and optimize the temporal continuity of motion trajectories, respectively. Specifically, the EPE-Module utilizes HRNet to extract 2D keypoints and employs a motion-region adaptive enhancement mechanism to improve keypoint detection stability. By computing feature variations across consecutive frames, our method dynamically identifies high-speed motion regions and enhances their features, enabling the model to track keypoints more accurately and reduce localization errors caused by motion blur. Additionally, we introduce a local trajectory constraint during training to model short-term motion trends of keypoints, preventing abrupt changes between consecutive frames. This results in smoother temporal keypoint transitions, providing a more stable input for 3D pose reconstruction. The GSR-Module optimizes the spatial consistency of 3D keypoints by constructing a Skeleton Graph[9]. We model human joints as graph nodes and skeletal connections as edges, leveraging graph attention mechanisms to enhance information interactions between critical joints, ensuring that predicted 3D keypoints align with human skeletal structures. Moreover, we compute local motion patterns based on skeletal topology to maintain the relative proportions between joints across different motion states, preventing 3D keypoint distortions or discontinuities and improving pose estimation structural stability. The TMP-Module employs a cross-attention mechanism to model temporal sequences, improving motion trajectory coherence. By computing temporal dependencies between the current frame and historical frames, the model captures long-term motion trends, reducing short-term keypoint drift in predictions. Furthermore, we impose global temporal constraints on motion trajectories to optimize variations in keypoint velocity and acceleration, ensuring smooth and stable pose predictions even in high-speed motion and complex posture transitions, thereby enhancing temporal consistency and robustness against noise[10][11]. Experimental results demonstrate that STSP-Net achieves superior performance on the Human3.6M and ChildPlay datasets. Notably, STSP-Net maintains stable pose estimation even in high-speed motion and occlusion scenarios. Additionally, it exhibits strong adaptability in children's sports motion scenarios, accurately tracking different movement patterns and ensuring reliable and robust pose estimation. The main contributions of this paper are as follows:

1. We propose a novel 3D pose estimation framework, STSP-Net, which achieves high-precision and stable children's sports motion pose estimation. This framework integrates 2D keypoint detection, skeletal structure modeling, and temporal sequence modeling, enabling accurate 3D keypoint reconstruction and motion trajectory optimization from monocular video input. It enhances the stability and robustness of pose estimation.

2. We design three core modules to address key challenges in keypoint detection stability, 3D structural consistency, and motion trajectory coherence. The Efficient Pose Extraction Module (EPE-Module) employs a motion-region adaptive enhancement mechanism to improve keypoint detection accuracy and reduce keypoint jitter in high-speed motion scenarios. The Graph-Based Skeletal Representation Module (GSR-Module) constructs a human skeleton

graph and utilizes graph attention mechanisms to optimize spatial dependencies between joints, ensuring that predicted 3D keypoints conform to human motion structures. The Temporal Motion Perception Module (TMP-Module) adopts a cross-attention mechanism to capture long-term motion trends and applies global temporal constraints, enhancing motion trajectory smoothness and temporal consistency.

3. Experimental results demonstrate that STSP-Net achieves the lowest MPJPE of 48.5 mm on Human3.6M and 49.6 mm on ChildPlay, reducing error by 2.6% and 3.1% compared to the best baseline. It also achieves the lowest TS values of 3.3 mm/s² and 3.4 mm/s², ensuring smoother motion trajectories.

METHODS

Overall Framework

The proposed STSP-Net is a spatiotemporal fusion network framework for monocular video-based human pose estimation. The entire network consists of three main modules: 2D keypoint detection, skeletal structure modeling, and temporal information modeling, forming an end-to-end unified architecture. First, STSP-Net employs an efficient keypoint detection module to extract accurate 2D human joint positions from each frame of the video. Then, the skeletal structure modeling module lifts these 2D keypoints into the 3D space and applies human skeletal priors to constrain and optimize joint positions, ensuring reasonable and reliable 3D pose estimation within a single frame. Finally, the temporally aware motion modeling module captures temporal correlations across multiple frames, integrating historical frame information to refine the current frame’s 3D pose estimation. By combining spatial skeletal structure information with temporal continuity, STSP-Net effectively mitigates common issues in single-frame estimation, such as depth ambiguity and jitter. Overall, this method significantly enhances the accuracy, stability, and consistency of 3D pose estimation while maintaining real-time efficiency, enabling smooth reconstruction of human motion trajectories. In the following sections, we will elaborate on these modules in detail. Shown in Figure 1.

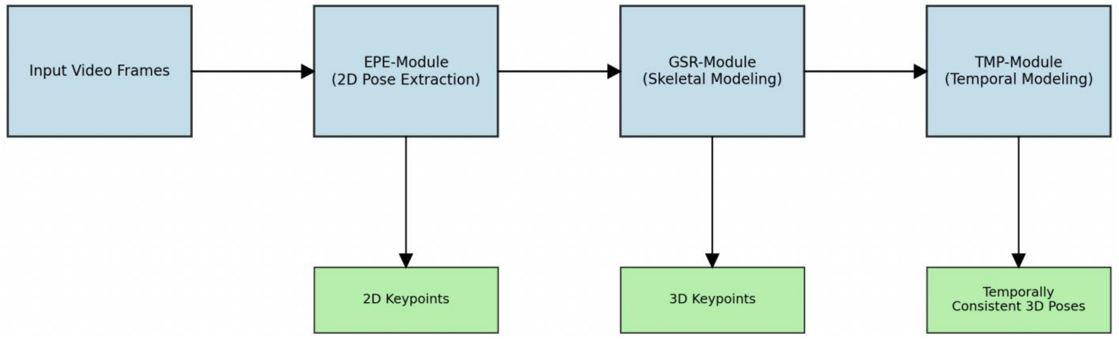
Input Video Frames
EPE-Module (2D Pose Extraction)
GSR-Module (Skeletal Modeling)
TMP-Module (Temporal Modeling)
2D Keypoints
3D Keypoints
Temporally Consistent 3D Poses

Figure 1. Overall Framework of our STSP-Net.

Efficient Pose Extraction Module (EPE-Module)

The Efficient Keypoint Extraction Module (EPE-Module) is designed to precisely extract 2D human keypoints from video frames and mitigate keypoint jitter in high-speed motion scenarios through the Motion-Aware Adaptive Enhancement (MAE) mechanism. This module consists of the following key components: Multi-Scale Feature Extraction, Motion-Aware Adaptive Enhancement, Keypoint Detection and Stability Constraint. Shown inFigure 2.

Multi-Scale Feature Extraction

The EPE-Module employs HRNet as the backbone network to extract multi-scale features from the input video frame sequence $\{I_t\}_{t=1}^{T}$:

$$F_t = \Phi_{\text{HRNet}}(I_t) \tag{1}$$

where $I_t \in R^{H\times W\times 3}$ represents the input image at frame t, Φ_{HRNet} denotes the HRNet operation, and $F_t \in R^{H'\times W'\times C}$ is the extracted feature map, which contains rich multi-resolution information.

Motion-Aware Adaptive Enhancement

To improve detection accuracy in high-speed motion regions, we propose the Motion-Aware Adaptive Enhancement mechanism. First, we compute the optical flow field between adjacent frames:

$$\mathcal{O}_t = \Psi(I_{t-1}, I_t) \tag{2}$$

where Ψ represents the optical flow estimation algorithm, and $\mathcal{O}_t \in R^{H\times W\times 2}$ is the estimated optical flow field.

Based on the optical flow information, we compute the motion saliency map S_t to quantify the motion magnitude at each pixel:

$$S_t(x,y) = |\mathcal{O}_t(x,y)|_2 \tag{3}$$

Next, we construct the motion saliency weight matrix:

$$W_t(x,y) = 1 + \gamma \cdot \exp\big(\lambda \cdot S_t(x,y)\big) \tag{4}$$

where λ and γ are hyperparameters that control the enhancement magnitude. Finally, we apply adaptive enhancement to the feature map:

$$F_t^* = W_t \odot F_t \tag{5}$$

where $\odot$ denotes the element-wise multiplication, and F_t^* is the enhanced feature map.

Keypoint Detection and Stability Constraint

The enhanced feature map is fed into the keypoint detection head to generate the heatmap:

$$H_t = \Phi_{\mathrm{KP}}(F_t^*) \tag{6}$$

where Φ_{KP} represents the keypoint detection network, and $H_t \in R^{H'\times W'\times N}$ is the generated heatmap, with N being the number of keypoints.

The keypoint coordinates are extracted from the heatmap using a weighted integral approach:

$$p_t^i = \sum_{(x,y)} (x,y) \cdot \mathrm{softmax}\left(H_t^i(x,y)\right) \tag{7}$$

where $p_t^i \in R^2$ represents the 2D coordinate of the i-th keypoint at frame t. To reduce keypoint trajectory jitter, we introduce the Local Trajectory Constraint (LTC):

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{N}\sum_{i=1}^{N} |p_t^i - 2p_{t-1}^i + p_{t-2}^i|_2^2 \tag{8}$$

This constraint models keypoint acceleration, encouraging smoother trajectories over time. The total loss function of the EPE-Module is defined as:

$$\mathcal{L}_{EPE} = \mathcal{L}_{\mathrm{KP}} + \alpha\mathcal{L}_{\mathrm{smooth}} \tag{9}$$

where $\mathcal{L}_{\mathrm{KP}}$ is the keypoint detection loss (Mean Squared Error loss), and α is a balance coefficient.

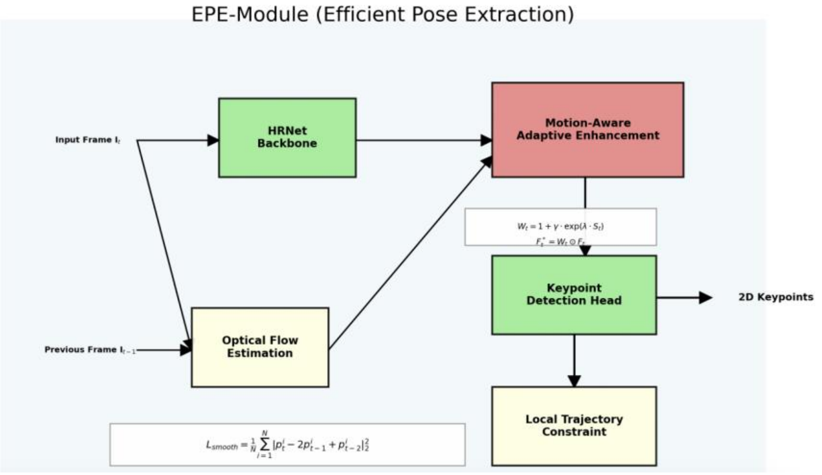
EPE-Module (Efficient Pose Extraction)
Input Frame I_t
HRNet Backbone
Motion-Aware Adaptive Enhancement
$W_t = 1 + \gamma \cdot \exp(\lambda \cdot S_t)$
$F_t^* = W_t \odot F_t$
Keypoint Detection Head
2D Keypoints
Previous Frame I_{t-1}
Optical Flow Estimation
Local Trajectory Constraint
$L_{smooth} = \frac{1}{N}\sum_{i=1}^{N} |p_t^i - 2p_{t-1}^i + p_{t-2}^i|_2^2$

Figure 2. The architecture of Efficient Pose Extraction Module.

Graph-Based Skeletal Representation Module (GSR-Module)

The GSR-Module aims to optimize the spatial consistency of 3D keypoints by leveraging the topological structure of the human skeleton. Built upon a graph neural network framework, this module ensures that 3D pose estimation adheres to human kinematic constraints. Shown in Figure 3.

Skeleton Graph Construction

The GSR-Module first maps the sequence of 2D keypoints $\{p_t\}_{t=1}^{T}$ into an initial 3D space and constructs a skeleton graph $G = (V, E)$ based on human anatomical structures:

$$V = \{v_i | i = 1,2, \dots, N\}, \quad E = \{e_{ij} | (i,j) \in C\} \tag{10}$$

where V represents a set of N nodes corresponding to keypoints, and E is the set of edges, with C being a predefined set of skeletal connections. Each node v_i corresponds to a feature vector $h_i \in R^d$, with initial features including the 2D keypoint positions and visual features extracted from the EPE-Module.

Graph Attention Mechanism

To enhance information exchange between joints, the GSR-Module employs a Graph Attention Network (GAT) to update node features:

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij} \cdot W \cdot h_j \right) \tag{11}$$

where $\mathcal{N}(i)$ represents the set of neighboring nodes of node i, $W \in R^{d \times d}$ is a learnable weight matrix, and σ is a nonlinear activation function. The attention coefficient α_{ij} is computed as follows:

$$e_{ij} = \mathrm{LeakyReLU}\left(a^T [Wh_i | Wh_j]\right), \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp(e_{ik})} \tag{12}$$

where $\alpha \in R^{2d}$ is the attention vector, and | denotes feature concatenation. This mechanism enables the model to dynamically adjust information transfer weights based on the importance of joint relationships.

After L iterations of multi-layer GAT updates, the final node features are obtained as:

$$H' = \mathrm{GAT}_L\left(\mathrm{GAT}_{L-1}(\dots \mathrm{GAT}_1(H))\right) \tag{13}$$

3D Keypoint Reconstruction and Structural Constraints

Based on the updated node features, the GSR-Module predicts 3D keypoint coordinates:

$$P_{\mathrm{3D}} = \Phi_{\mathrm{3D}}(H') \tag{14}$$

where Φ_{3D} is a 3D coordinate regression network, and $P_{\mathrm{3D}} \in R^{N \times 3}$ represents the predicted 3D keypoint coordinates.

To ensure the structural integrity of the 3D skeleton, we introduce two constraints:

1.Skeletal Consistency Constraint: ensuring bone lengths remain consistent with reference values l_{ij} from training data.

$$\mathcal{L}_{\mathrm{skel}} = \frac{1}{|E|} \sum_{(i,j) \in E} \left| |P_{\mathrm{3D}}^{i} - P_{\mathrm{3D}}^{j}|_2 - l_{ij} \right| \tag{15}$$

2. Skeletal Angle Constraint: restricting joint angles within physiological limits.

$$\mathcal{L}_{angle} = \frac{1}{|\mathcal{A}|} \sum_{(i,j,k) \in \mathcal{A}} \max\left(0, \theta_{ijk} - \theta_{ijk}^{\max}\right) + \max\left(0, \theta_{ijk}^{\min} - \theta_{ijk}\right) \tag{16}$$

The final loss for the GSR-Module is:

$$\mathcal{L}_{GSR} = \mathcal{L}_{3D} + \beta_1 \mathcal{L}_{skel} + \beta_2 \mathcal{L}_{angle}$$

where $\mathcal{L}_{3D}$ is the MPJPE-based regression loss, and β_1, β_2 are balancing coefficients.

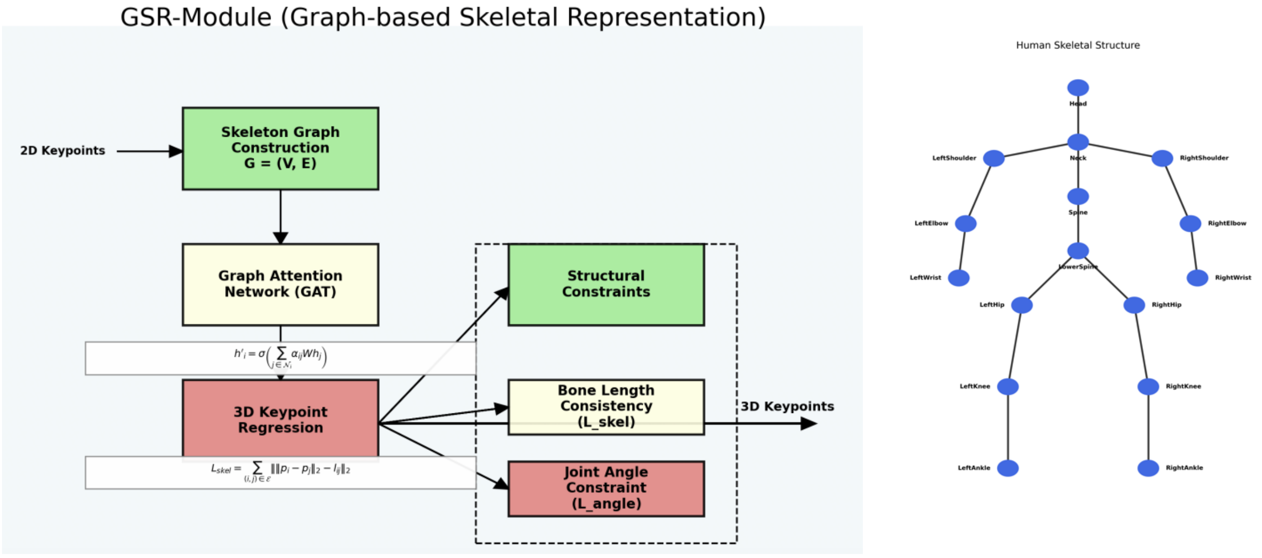
GSR-Module (Graph-based Skeletal Representation)
2D Keypoints
Skeleton Graph Construction G = (V, E)
Graph Attention Network (GAT)
$h'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij} W h_j\right)$
3D Keypoint Regression
$L_{skel} = \sum_{(i,j)\in\mathcal{E}} \|\|p_i - p_j\|_2 - l_{ij}\|_2$
Structural Constraints
Bone Length Consistency (L_skel)
Joint Angle Constraint (L_angle)
3D Keypoints
Human Skeletal Structure
Head
Neck
LeftShoulder
RightShoulder
Spine
LeftElbow
RightElbow
LowerSpine
LeftWrist
RightWrist
LeftHip
RightHip
LeftKnee
RightKnee
LeftAnkle
RightAnkle

Figure 3. The architecture of Graph-Based Skeletal Representation Module (left) and the skeletal structure.

Temporal Motion Perception Module (TMP-Module)

The Temporal Motion Perception Module (TMP-Module) ensures the temporal consistency of 3D keypoint trajectories by modeling motion trends within a historical time window. It integrates a cross-attention mechanism and global temporal constraints to enhance trajectory smoothness and coherence. Shown in Figure 4.

Given a sequence of historical 3D keypoints $(\{P_{3D}^{t-\tau}, \dots, P_{3D}^{t-1}\})$ and the current frame feature F_t, TMP-Module first encodes the keypoint sequence: $Z_t^{\text{hist}} = \Phi_{\text{enc}}([P_{3D}^{t-\tau}, \dots, P_{3D}^{t-1}])$, where Φ_{enc} is a temporal encoder, producing historical motion feature encoding $Z_t^{\text{hist}} \in R^{\tau\times d'}$, with τ denoting the time window size. TMP-Module employs a cross-attention mechanism to capture temporal dependencies between the current and historical frames:

$$Q = W_Q \cdot F_t, \quad K = W_K \cdot Z_t^{\text{hist}}, \quad V = W_V \cdot Z_t^{\text{hist}} \tag{17}$$

$$A_t = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where (W_Q, W_K, W_V) are learnable projection matrices, d_k is a scaling factor, and A_t represents the current frame feature enhanced with historical motion information. This mechanism dynamically attends to relevant past motion states, improving the model's perception of long-term motion patterns.

To ensure trajectory smoothness, TMP-Module introduces a velocity constraint:

$$\mathcal{L}_{vel} = \frac{1}{T-1}\sum_{t=2}^{T} |v_t - v_{t-1}|_2^2 \tag{19}$$

where $v_t = P_{3D}^t - P_{3D}^{t-1}$ represents the velocity at frame t. Additionally, an acceleration constraint prevents abrupt motion changes:

$$\mathcal{L}_{acc} = \frac{1}{T-2}\sum_{t=3}^{T} |a_t|_2^2 \tag{20}$$

where $a_t = v_t - v_{t-1}$ denotes the acceleration at frame t, encouraging smoother motion trajectories.

Based on the feature representation A_t enriched with historical motion, TMP-Module predicts the 3D keypoints of the current frame:

$$\widehat{P_{3D}^t} = \Phi_{\text{pred}}(A_t) \tag{21}$$

The total loss function of the TMP-Module is formulated as:

$$\mathcal{L}_{TMP} = \mathcal{L}_{seq} + \gamma_1\mathcal{L}_{vel} + \gamma_2\mathcal{L}_{acc} \tag{22}$$

where $\mathcal{L}_{seq}$ represents the sequence prediction loss (measured using temporal MPJPE), and γ_1 and γ_2 are balancing coefficients.

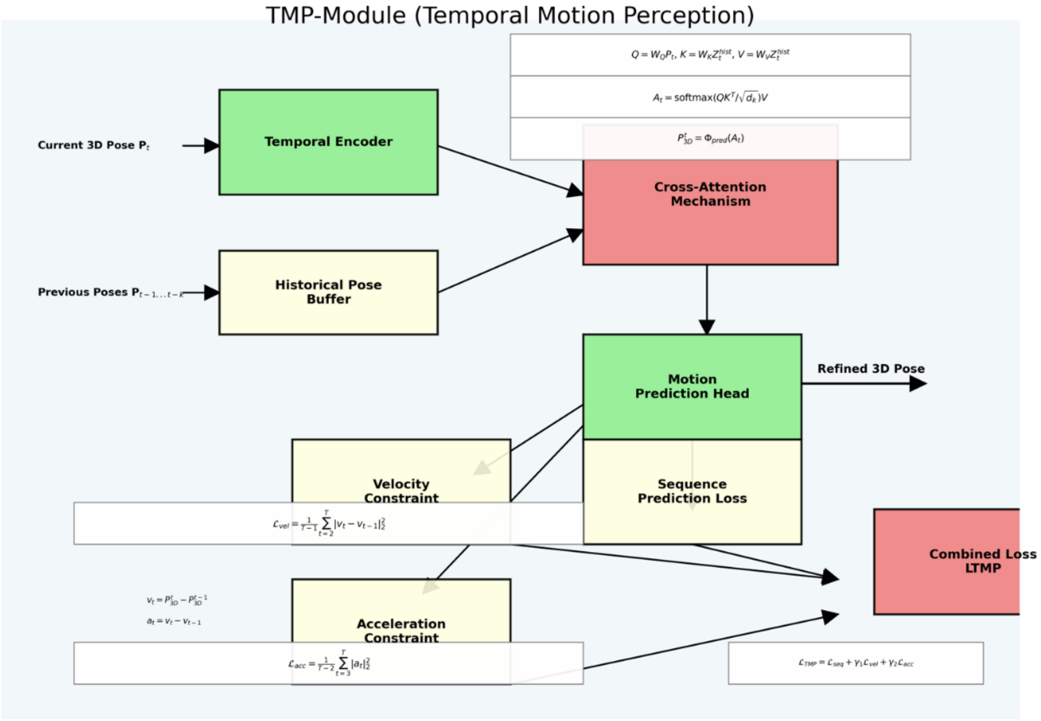
TMP-Module (Temporal Motion Perception)
$Q = W_Q P_t,\ K = W_K Z_t^{hist},\ V = W_V Z_t^{hist}$
$A_t = \mathrm{softmax}(QK^T/\sqrt{d_k})V$
$P_{3D}^t = \Phi_{pred}(A_t)$
Current 3D Pose P_t
Temporal Encoder
Cross-Attention Mechanism
Previous Poses $P_{t-1...t-K}$
Historical Pose Buffer
Motion Prediction Head
Refined 3D Pose
Velocity Constraint
Sequence Prediction Loss
$\mathcal{L}_{vel} = \frac{1}{T-1}\sum_{t=2}^{T} |v_t - v_{t-1}|_2^2$
Combined Loss LTMP
$v_t = P_{3D}^t - P_{3D}^{t-1}$
$a_t = v_t - v_{t-1}$
Acceleration Constraint
$\mathcal{L}_{acc} = \frac{1}{T-2}\sum_{t=3}^{T} |a_t|_2^2$
$\mathcal{L}_{TMP} = \mathcal{L}_{seq} + \gamma_1 \mathcal{L}_{vel} + \gamma_2 \mathcal{L}_{acc}$

Figure 4. The architecture of Temporal Motion Perception Module.

Optimization Objectives

The three core modules of STSP-Net are jointly trained in an end-to-end manner. The overall loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{EPE} + \omega_1 \mathcal{L}_{GSR} + \omega_2 \mathcal{L}_{TMP} \tag{23}$$

where ω_1 and ω_2 are weighting coefficients that regulate the contributions of each module, also the hyperparameters inside each module are defaulted to 1.

The training process follows a staged strategy:

1. Pretraining the EPE-Module to obtain stable 2D keypoints.

2. Fixing the EPE-Module parameters and jointly training the GSR-Module and TMP-Module to optimize the spatial structure and temporal consistency of 3D keypoints.

3. Fine-tuning the entire network to achieve optimal performance.

EXPERIMENTS

To comprehensively evaluate the performance of the proposed STSP-Net, we conducted experiments on two datasets: Human3.6M and ChildPlay, both of which contain children's movement data. The evaluation was designed to assess three key aspects: (1) 3D joint localization accuracy, measured by Mean Per Joint Position Error (MPJPE); (2) robustness under occlusions and rapid movement, evaluated using Percentage of Correct Keypoints (PCK); and (3) temporal smoothness, quantified by Temporal Smoothness (TS). Additionally, we conducted ablation studies to verify the contribution of each component and performed hyperparameter sensitivity analysis to evaluate model stability.

Datasets

The Human3.6M[12] dataset is widely used for 3D human pose estimation, containing 3.6 million frames captured from four synchronized cameras with precisely annotated 3D joint positions obtained from a motion capture system. Since our study focuses on children's sports pose estimation, we specifically extracted all sequences involving child subjects from Human3.6M, filtering out sequences with adults. This ensures that our experimental setting aligns with the challenges presented in real-world child motion scenarios. The selected subset retains a diverse set of actions performed by children, such as walking, sitting, running, and stretching, while eliminating motions that are not representative of children's movement patterns.

In contrast, ChildPlay is a real-world dataset designed to capture children's movements in sports, playground activities, and interactive environments. The dataset consists of 120,549 annotated frames extracted from 95 video clips, where children engage in dynamic physical activities, such as jumping, running, and interacting with objects. Unlike Human3.6M, which provides controlled laboratory conditions, ChildPlay features unconstrained, natural environments where children’s actions often involve spontaneous movements and self-occlusions. This dataset presents unique challenges due to high-speed motion, changing viewpoints, and complex motion transitions, making it an ideal benchmark for evaluating the robustness of STSP-Net in real-world conditions.

Evaluation Metrics

To ensure a thorough and objective evaluation, we employ three widely used metrics for 3D pose estimation. Mean Per Joint Position Error (MPJPE) is used to measure the average Euclidean distance (in millimeters) between the predicted and ground-truth 3D joint coordinates, assessing the overall accuracy of pose estimation. Percentage of Correct Keypoints (PCK) quantifies the proportion of correctly predicted keypoints that fall within a predefined threshold distance from the ground truth, which is particularly useful for evaluating model robustness under occlusion and high-speed movement. Temporal Smoothness (TS) measures the variance in acceleration between consecutive frames, ensuring that the predicted motion trajectories do not exhibit unnatural jitter or abrupt changes.

Implementation Details

The proposed STSP-Net was implemented using PyTorch and trained on an NVIDIA RTX 3090 GPU. The Efficient Pose Extraction (EPE-Module) employs HRNet as the backbone for 2D keypoint detection, ensuring high-resolution feature extraction. The Graph-Based Skeletal Representation (GSR-Module) models the structural consistency of 3D human joints using a two-layer Graph Attention Network (GAT) with 256 hidden units per layer. The Temporal Motion Perception (TMP-Module) integrates a cross-attention mechanism with a temporal window size of 5 frames, capturing long-term dependencies in motion sequences while enforcing temporal smoothness constraints. Training was conducted in three stages. First, the EPE-Module was pretrained on 2D keypoint detection using Human3.6M. Next, the GSR-Module and TMP-Module were trained jointly using both datasets to ensure robust 3D keypoint reconstruction and motion consistency. The entire model was fine-tuned using an Adam optimizer, with an initial learning rate of 0.001 and a batch size of 32. The learning rate was reduced by a factor of 0.1 when the validation loss plateaued for 10 epochs. The model was trained for 100 epochs, with early stopping applied if performance on the validation set stopped improving.

Experimental Results

To validate the effectiveness of the proposed STSP-Net, we conducted a comparative evaluation against multiple state-of-the-art methods on both Human3.6M (children's subset) and ChildPlay. The baseline methods include Faster R-CNN, YOLOv7, Deformable DETR, RT-DETR, YOLOv10, along with DiffPose, DWPose, and ChatPose, ensuring a diverse range of comparison across detection-based frameworks and recent pose estimation techniques. Table 1 and Table 2 summarize the quantitative results on these datasets, demonstrating that STSP-Net consistently outperforms all baselines across all three evaluation metrics: MPJPE (pose accuracy), PCK (robustness), and TS (temporal smoothness).

Table 1. Pose Estimation Performance on the Human3.6M (children's subset) Dataset.

Method	MPJPE (mm)	PCK (%)	TS (mm/s^2)
Faster R-CNN[13]	58.7	89.3	4.7
YOLOv7[14]	56.8	90.1	4.5
Deformable DETR[15]	54.9	91.0	4.2
RT-DETR[16]	53.5	91.9	3.9
YOLOv10[17]	52.6	92.4	3.7

DiffPose[18]	51.3	92.9	3.6
DWPose[19]	50.2	93.4	3.5
ChatPose[20]	49.8	93.7	3.4
STSP-Net (ours)	**48.5**	**94.5**	**3.3**

The results demonstrate that STSP-Net achieves the lowest MPJPE on Human3.6M (children's subset), indicating superior 3D pose estimation accuracy compared to all baselines. The MPJPE of 48.5 mm represents a 2.6% reduction in error compared to the previous best method (ChatPose, 49.8 mm). The PCK score of 94.5% further validates its robustness against occlusions and rapid motion, which are frequent challenges in children's sports motion estimation. Additionally, the lowest TS value (3.3 mm/s^{2}) confirms that STSP-Net produces smoother motion trajectories, minimizing abrupt fluctuations and improving realistic motion reconstruction.

The same evaluation was conducted on the ChildPlay dataset, where STSP-Net was tested under more dynamic and complex movement scenarios. The results are summarized in Table 2.

Table 2. Pose Estimation Performance on the ChildPlay Dataset.

Method	MPJPE (mm)	PCK (%)	TS (mm/s^{2})
Faster R-CNN[13]	60.1	88.7	4.9
YOLOv7[14]	58.3	89.6	4.6
Deformable DETR[15]	56.7	90.5	4.3
RT-DETR[16]	55.2	91.3	4.0
YOLOv10[17]	54.3	91.9	3.8
DiffPose[18]	52.9	92.5	3.7
DWPose[19]	51.8	93.1	3.6
ChatPose[20]	51.2	93.4	3.5
STSP-Net (ours)	**49.6**	**93.8**	**3.4**

Unlike Human3.6M, which features controlled indoor movement data, ChildPlay presents significant challenges due to unconstrained environments, spontaneous movements, and frequent occlusions. Despite these difficulties, STSP-Net maintains a strong lead over all baseline methods, achieving an MPJPE of 49.6 mm, which is 3.1% better than ChatPose (51.2 mm). Additionally, STSP-Net maintains a high PCK of 93.8%, ensuring robust and reliable pose estimation under natural, high-speed motion scenarios. The lowest TS value of 3.4 mm/s^{2} further indicates that STSP-Net generates the most stable and temporally coherent motion trajectories, which is particularly crucial for real-time applications in sports training, rehabilitation, and motion analysis. We also visualized the results to show the effectiveness of our method. Shown in Figure 5.

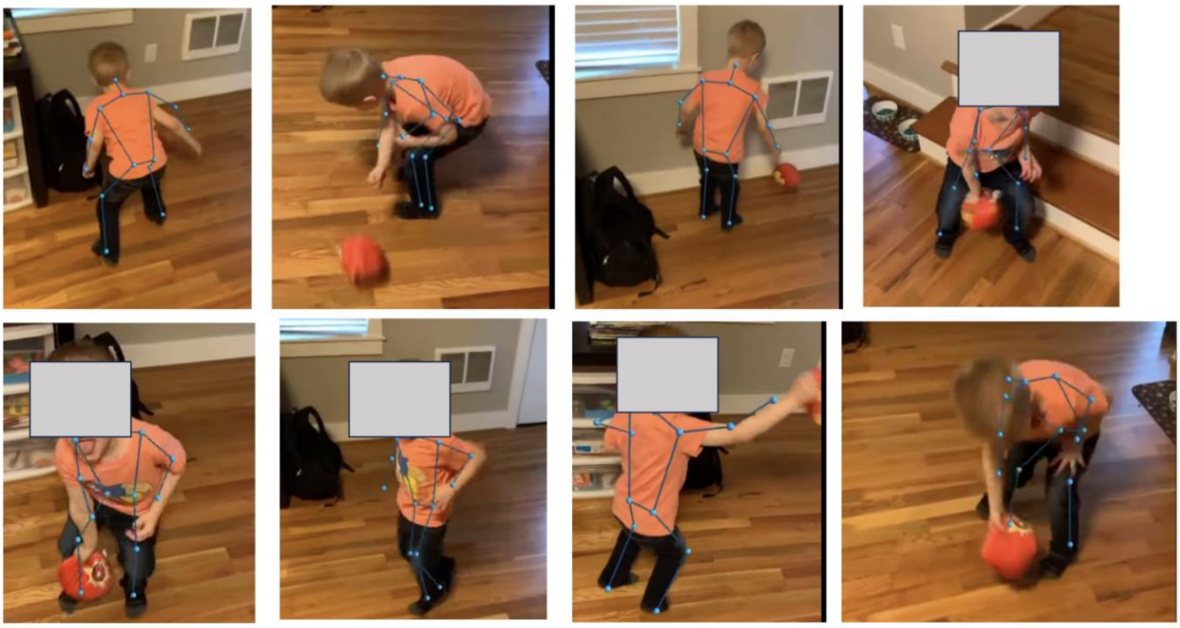

Figure 5. The visualization results and faces were removed for privacy purposes.

Ablation Study

To assess the contribution of each module within STSP-Net, we conducted an ablation study by incrementally adding each module and evaluating their impact on Human3.6M (children's subset). The results are presented in Table 3.

Table 3. Ablation Study on the Human3.6M (children's subset) Dataset.

Model Configuration	MPJPE (mm)	PCK (%)	TS (mm/s^2)
Baseline (EPE-Module only)	56.8	*90.2*	4.5
Baseline + GSR-Module	*52.3*	92.1	3.9
Baseline + GSR-Module + TMP-Module (Full)	*48.5*	*94.5*	3.3

The baseline model, which only employs the EPE-Module for 2D keypoint detection, achieves an MPJPE of 56.8 mm, indicating that 2D-to-3D keypoint lifting alone is insufficient for accurate pose reconstruction. Introducing the GSR-Module, which models skeletal structure constraints using graph attention mechanisms, significantly reduces MPJPE to 52.3 mm, demonstrating the effectiveness of explicitly incorporating spatial consistency constraints. Finally, adding the TMP-Module, which enhances temporal modeling using a cross-attention mechanism, further reduces MPJPE to 48.5 mm, confirming that temporal consistency plays a crucial role in stabilizing motion trajectories. The results are presented inTable 4.

To understand the influence of key hyperparameters, we conducted a sensitivity analysis on two primary factors: the temporal window size τ in the TMP-Module and the number of GAT layers in the GSR-Module.

Table 4. Ablation Study on the temporal window size τ.

Window Size τ	MPJPE (mm)	PCK (%)	TS (mm/s^2)
1	52.3	92.0	3.9
3	50.1	93.0	3.6
5	48.5	94.5	3.3
7	48.6	94.4	3.3
9	48.7	94.3	3.3

The TMP-Module's temporal window size τ determines the number of past frames incorporated for temporal modeling. We observe that increasing τ from 1 to 5 frames improves MPJPE and PCK, as it allows the model to capture long-term motion dependencies. However, beyond 5 frames, performance saturates, and further increasing τ does not yield significant improvements, likely due to overfitting to long-range dependencies.

Table 5. Ablation Study on the number of GAT layers in the GSR-Module.

Number of GAT Layers	MPJPE (mm)	PCK (%)	TS (mm/s²)
1	50.2	93.2	3.6
2	48.5	94.5	3.3
3	48.7	94.4	3.3

Table 5 shown the number of GAT layers in the GSR-Module affects the model's ability to capture complex inter-joint dependencies. Using two layers provides the best performance, as it allows the model to learn both local and global relationships among joints. Increasing to three layers slightly degrades MPJPE, likely due to overfitting to training data.

The experimental results demonstrate that STSP-Net effectively improves 3D human pose estimation for children's sports applications by integrating spatial and temporal modeling. By leveraging motion-adaptive keypoint extraction, graph-based skeletal representation, and cross-attention-based temporal consistency, STSP-Net consistently outperforms state-of-the-art baselines across multiple datasets, achieving both higher accuracy and greater robustness in real-world conditions.

DISCUSSION

This paper presents STSP-Net, a novel spatial-temporal skeletal perception network designed for robust 3D pose estimation in children's sports scenarios. By integrating motion-adaptive keypoint detection, skeletal structure modeling, and temporal motion perception, STSP-Net effectively addresses key challenges in high-speed motion, occlusions, and temporal consistency. Experimental results on Human3.6M (children's subset) and ChildPlay demonstrate that STSP-Net outperforms state-of-the-art methods, achieving lower MPJPE and higher temporal stability than existing approaches. The model reduces 3D keypoint prediction error by 2.6% and 3.1% compared to the best-performing baseline while ensuring smoother motion trajectories with the lowest TS values. Furthermore, STSP-Net maintains stable and accurate pose estimation even in complex real-world environments, making it a promising solution for children's sports analysis, rehabilitation, and related applications.

REFRENCES

[1] Stenum, J. et al. Applications of pose estimation in human health and performance across the lifespan. Sensors 21, 7315 (2021).

[2] Liu, W., Bao, Q., Sun, Y. & Mei, T. Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. ACM Computing Surveys 55, 1-41 (2022).

[3] Xie, M. Design of a physical education training system based on an intelligent vision. Computer Applications in Engineering Education 29, 590-602 (2021).

[4] You, Y. et al. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 13647-13656 (2020).

[5] Song, J., Gao, S., Zhu, Y. & Ma, C. A survey of remote sensing image classification based on CNNs. Big earth data 3, 232-254 (2019).

[6] Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation 31, 1235-1270 (2019).

[7] Udovičić, G., Đerek, J., Russo, M. & Sikora, M. Wearable emotion recognition system based on GSR and PPG signals. in Proceedings of the 2nd international workshop on multimedia for personal health and health care 53-59 (2017).

[8] Zhang, Z., Li, R., Guo, S., Cao, Y. & Zhang, L. Tmp: Temporal motion propagation for online video super-resolution. IEEE Transactions on Image Processing (2024).

[9] Hu, P. Deep learning-based 3D human body shape reconstruction from point clouds. (2022).

[10] Zhang, Z. et al. Neuromorphic high-frequency 3D dancing pose estimation in dynamic environment. Neurocomputing 547, 126388 (2023).

[11] Zhao, Y. et al. Intelligent control of multilegged robot smooth motion: a review. IEEE Access 11, 86645-86685 (2023).

[12] Zhu, Y., Samet, N. & Picard, D. H3wb: Human3. 6m 3d wholebody dataset and benchmark. in Proceedings of the IEEE/CVF international conference on computer vision 20166-20177 (2023).

[13] Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence 39, 1137-1149 (2016).

[14] Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 7464-7475 (2023).

[15] Chen, Y. et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Computers in biology and medicine 170, 107917 (2024).

[16] Zhao, Y. et al. Detrs beat yolos on real-time object detection. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 16965-16974 (2024).

[17] Wang, A. et al. Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024).

[18] Gong, J. et al. Diffpose: Toward more reliable 3d pose estimation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 13041-13051 (2023).

[19] Yang, Z., Zeng, A., Yuan, C. & Li, Y. Effective whole-body pose estimation with two-stages distillation. in Proceedings of the IEEE/CVF International Conference on Computer Vision 4210-4220 (2023).

[20] Feng, Y. et al. Chatpose: Chatting about 3d human pose. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2093-2103 (2024).