Machine Vision and Applications (2026) 37:82
https://doi.org/10.1007/s00138-026-01852-7

RESEARCH

SBAHGNet:3D human pose estimation via skeleton-biased attention and high-frequency enhanced graph convolution

Yu Wang[1] · Jiaqiu Ai[1] · Xinyu Sun[1] · Yong Zhang[1] · Jinyang Huang[1]

Received: 8 January 2026 / Revised: 30 April 2026 / Accepted: 24 May 2026

Abstract

Monocular 3D human pose estimation is challenged by depth ambiguity and complex articulation, which complicate feature modeling and demand robust spatio-temporal representations. Although existing methods have advanced spatio-temporal modeling, limitations remain: graph convolutional network (GCN) exhibits low-pass behavior that, as depth increases, attenuates high-frequency geometric details in joint trajectories and thus degrades depth accuracy; and standard self-attention does not explicitly encode skeletal topology, resulting in indirect modeling of bone connectivity. To address these issues, we propose SBAHGNet, a dual-branch spatio-temporal feature-fusion network. In the GCN branch, a Multi-Scale High-Frequency Enhancement (MSHFE) module—applied after feature aggregation-recovers high-frequency geometric cues lost to GCN smoothing, improving fine-grained depth representation. In the attention branch, a Skeletal-Biased Attention (SBA) module injects a learnable skeletal bias into spatial attention to explicitly encode skeletal topology and strengthen structural modeling. Complementary features from both branches are adaptively fused for final 3D pose regression. Extensive experiments on Human3.6 M and MPI-INF-3DHP validate our approach. With detected 2D keypoints, SBAHGNet attains 37.24 mm MPJPE (P1) and 31.57 mm PA-MPJPE (P2) on Human3.6 M (12.38 mm with ground-truth 2D), and 13.83 mm MPJPE, 99.02% PCK@150 mm, and 88.22 AUC on MPI-INF-3DHP. With only 18.3 M parameters, the model achieves a favorable accuracy–efficiency trade-off and outperforms many comparable methods.

Keywords Monocular 3D human pose estimation · Spatio-temporal fusion · Graph convolutional network · Skeletal-Biased Attention · High-frequency enhancement

1 Introduction

Monocular 3D human pose estimation aims to recover the three-dimensional coordinates of human keypoints from a single RGB image or video sequence, serving as a foundational task for understanding human motion and behavior [1, 2].Owing to its reliance on inexpensive standard RGB cameras for capturing 3D poses, monocular pose estimation offers low cost and high practicality, finding widespread applications in action recognition [3], intelligent surveillance [4], human-computer interaction [5], augmented reality [6], and virtual reality [7]. However, the inherent depth ambiguity in monocular inputs (where poses at varying depths may yield similar 2D projections), coupled with motion blur from rapid movements, occlusions, and noise in 2D detections, poses significant challenges to effective spatio-temporal feature modeling [8, 9].

In engineering practice, the 2D-to-3D lifting paradigm, which first detects 2D keypoints and subsequently regresses 3D coordinates, has emerged as the dominant approach [10–12]. This paradigm capitalizes on mature 2D detectors while substantially reducing annotation and training costs. Its core challenge lies in recovering accurate and skeleton-consistent 3D structures from noisy and potentially incomplete

✉ Jiaqiu Ai
aijiaqiu1985@hfut.edu.cn

Yu Wang
2024170780@mail.hfut.edu.cn

Xinyu Sun
2025010057@mail.hfut.edu.cn

Yong Zhang
yongzhang@hfut.edu.cn

Jinyang Huang
hjy@hfut.edu.cn

[1] School of Computer and Information, Hefei University of Technology, Hefei 230009, Anhui Province, China

Published online: 02 June 2026

2D keypoint sequences, imposing stringent requirements on the expressive capacity and robustness of spatio-temporal modeling strategies. Early works primarily relied on Temporal Convolutional Network (TCN) for sequence modeling. For instance, VideoPose3D [10] proposed by Pavllo et al. employed dilated temporal convolutions to achieve large receptive fields and incorporated semi-supervised 2D-3D reprojection to exploit unlabeled videos, establishing a widely adopted benchmark. Although TCN is computationally efficient and straightforward to implement, it exhibits clear limitations in capturing extremely long-range dependencies and complex cross-joint interactions [13]. To address these limitations, researchers have pursued two complementary directions: introducing Graph Convolutional Network (GCN) to explicitly encode skeletal topology and reinforce local kinematic priors on one hand, and adopting attention-based mechanisms to model global dependencies across joints and frames on the other [14–17].

GCN effectively captures structured inter-joint relationships through neighborhood aggregation (e.g., SemGCN [15] demonstrated high parameter efficiency and robustness), significantly enhancing skeleton consistency and local geometric representation. However, GCN inherently possesses low-pass filtering characteristics [18–20]. As network depth increases, features progressively homogenize, attenuating high-frequency details such as rapid joint trajectories and local geometric variations, thereby impairing depth recovery and subtle motion estimation. Subsequent studies have proposed multi-scale designs, spectral decompositions, and trainable bandpass filtering strategies to compensate for lost high-frequency components [21, 22]. To tackle this issue, we introduce a Multi-Scale High-Frequency Enhancement (MSHFE) module in the GCN branch. Applied after feature aggregation, this module selectively restores suppressed multi-scale high-frequency components, recovering geometric details in joint motions and improving depth estimation accuracy.

Meanwhile, attention-based methods excel in global spatio-temporal modeling by directly capturing long-range interactions between arbitrary joints and frames [23–26]. A representative work, PoseFormer [27], pioneered the separation of spatial and temporal attention modules to encode intra-frame joint relationships and inter-frame motion dependencies, respectively. Follow-up approaches such as MixSTE [16] and PoseFormerV2 [28] further expand receptive fields and enhance robustness through refined token designs or frequency-domain representations (e.g., preserving critical high-frequency coefficients in Discrete Cosine Transform). Despite the significant advantages of attention mechanisms in long-range dependency modeling, they typically process flattened joint sequences and lack explicit encoding of physical skeletal connections, leading to degraded skeleton consistency under noisy inputs [29, 30]. To address this limitation, we propose a Skeletal-Biased Attention (SBA) module that injects learnable skeletal topological bias into the spatial attention distribution with extremely low parameter overhead, explicitly guiding the attention mechanism toward anatomically relevant joint pairs and thereby enhancing local structural consistency.

Inspired by the complementary strengths of GCN's local structural modeling and attention mechanisms' global dependency capture, we propose SBAHGNet, a lightweight dual-branch spatio-temporal feature-fusion network designed to jointly achieve explicit skeletal topology injection, global long-range modeling, and high-frequency detail compensation. The graph branch incorporates the MSHFE module to recover high-frequency geometric cues lost due to GCN smoothing. The attention branch integrates the SBA module, which injects skeletal topological bias into the spatial attention distribution with negligible parameter overhead. Complementary features from both branches are adaptively fused for final 3D pose regression.

Our main contributions are as follows:

1. (1) We design the MSHFE module, which effectively compensates for high-frequency losses in the GCN branch, recovering geometric details in joint motions and enhancing depth estimation accuracy.
2. (2) We propose the SBA module, which injects skeletal topological bias into the spatial attention distribution with extremely low parameter overhead (376 additional parameters), improving local structural consistency.
3. (3) Through extensive experiments on the Human3.6 M and MPI-INF-3DHP datasets, we demonstrate that SBAHGNet achieves highly competitive performance among existing monocular 3D pose estimation methods.

2 Related work

Current research on monocular 3D human pose estimation primarily focuses on recovering 3D joint positions from single RGB images or video sequences, a task complicated by depth ambiguity, nonlinear motion, occlusion, and rapid movements [8, 9]. Recent approaches predominantly adopt the 2D-to-3D lifting paradigm, where 2D keypoints are first detected and then lifted to 3D via spatio-temporal modeling [11, 12]. Existing methods can be grouped by their core modeling components into four main categories.

2.1 Temporal modeling based on temporal convolution and recurrent operators

In video-based monocular 3D pose estimation, temporal information from 2D keypoint sequences is exploited to improve inter-frame consistency and robustness. Early methods employ recurrent networks, such as sequence-to-sequence LSTMs [31], to model temporal joint evolution and reduce jitter. Some real-time systems integrate CNN-based regression with skeletal fitting or temporal filtering for online stability, exemplified by VNect [32]. More recently, dilated temporal convolutional networks (TCNs) gain prominence due to their large receptive fields and parallelism, with VideoPose3D introducing semi-supervised back-projection to utilize unlabeled data. Recent works further explore the potential of temporal modules in modeling motion dynamics and global consistency; for example, You et al. [33] propose PMCE, a dual-stream "co-evolution" architecture where one branch lifts 2D joint sequences to the mid-frame 3D pose and the other aggregates cross-time image features using a temporal convolutional network; Zheng et al. [34] introduce the Retentive Network (RetNet), leveraging a large window of past frames and a few future frames to capture long-range.

dependencies, with a non-causal variant (NC-RetNet) and a knowledge-transfer training scheme; Hsu and Jang [35] utilize an RNN to predict bone lengths over the entire sequence and adjust 3D poses accordingly to enforce physical consistency. These approaches offer simple architectures, efficient training/inference, and strong deployability, but often face limitations in explicitly encoding skeletal topological priors and recovering high-frequency geometric details from rapid local depth variations.

2.2 Skeletal topology modeling based on graph convolutional network

To incorporate human body priors, graph-based methods represent joints and bones as graphs and apply GCN for spatial aggregation of dependencies. ST-GCN [36] pioneers spatio-temporal graph convolution in skeleton-based tasks, while SemGCN [15] introduces learnable or semantically guided adjacency matrices for improved accuracy under constrained parameters. However, standard GCN exhibits low-pass filtering characteristics, causing over-smoothing and loss of high-frequency motion with deeper layers. Subsequent works enhance the capture of pose structure by refining graph convolution operations; for instance, Azizi et al. [37] propose MöbiusGCN, which uses Möbius transformations in the spectral domain to explicitly model inter-joint rotations, achieving state-of-the-art accuracy with drastically fewer parameters (only 0.042 M in its lightest version); Zhang [38] introduces GroupGCN, decoupling shared aggregation into group convolutions with independent adjacency kernels per feature group and cross-group interaction; Yu et al. [13] present GLA-GCN, an adaptive global-local architecture with one branch aggregating spatio-temporal features over the entire skeleton graph and another refining per-joint features via independently connected layers. These methods demonstrate that dynamically learning or explicitly encoding skeletal graph structures (even geometric transformations) enhances GCN performance in topological prior encoding, but can lead to loss of high-frequency geometric details and may not fully integrate long-range global dependencies under complex motions.

2.3 Attention-based global spatio-temporal modeling

Attention mechanisms have emerged as powerful tools for capturing long-range dependencies, making them particularly suitable for global spatio-temporal modeling in video sequences. These methods directly model interactions between arbitrary joints and frames, thereby overcoming the inherent limitations of convolutional approaches in terms of local receptive fields and enabling effective capture of global dependencies over extended temporal ranges. PoseFormer [27] represents one of the pioneering works, being the first to decouple spatial and temporal attention modules to separately encode intra-frame joint relationships and inter-frame motion dependencies. This approach better captures cross-frame global spatio-temporal relationships and has provided an important foundation for subsequent research. Numerous follow-up studies have further refined attention-based frameworks to address challenges such as depth ambiguity and long-sequence modeling. For example, MHFormer [39] employs a multi-hypothesis generation mechanism to improve pose estimation accuracy while enhancing model efficiency through alternating spatio-temporal blocks; MixSTE [16] incorporates frequency-domain representations and optimizes token designs to expand receptive fields and improve robustness; STCFormer [40] adopts Spatio-Temporal Criss-Cross Attention, achieving global interactions with sub-quadratic complexity while integrating local convolutions to provide richer contextual information. Beyond purely attention-based designs, several studies have explored attention variants combined with Graph Convolutional Networks or convolutional backbones to enhance the efficiency of global modeling. For instance, GAST-Net [41] integrates graph attention with spatio-temporal convolutions, enabling adaptive joint weighting to capture global spatial relationships; the SaEGC-Net [42] proposes a simplified spatio-temporal attention module (SST-Att) embedded within a GCN framework, effectively modeling

long-range dependencies between non-adjacent joints while avoiding the quadratic complexity of conventional self-attention mechanisms. Despite the significant advantages of attention-based methods in global dependency modeling, they typically process flattened joint sequences or rely on implicit learning, resulting in a lack of explicit encoding of physical skeletal connections. This limitation can impair the model's understanding of human skeletal structure, particularly when handling rapid motions or complex actions. Furthermore, the absence of explicit structural awareness may hinder the effective preservation of high-frequency geometric details arising from rapid local movements, thereby affecting the accuracy of depth estimation and subtle motion capture.

2.4 Hybrid architecture and spectral/frequency-domain enhancement

Recent efforts combine paradigms to address inductive biases and leverage frequency representations for efficiency. For instance, Zhao et al. [28] propose PoseFormerV2, converting long joint sequences to the frequency domain via Discrete Cosine Transform, using few low-frequency coefficients to expand receptive fields and substantially reduce computation, fusing time- and frequency-domain features for better speed-accuracy trade-off and noise robustness; Lin et al. [43] introduce AMPose, alternately stacking Transformer and GCN layers to jointly encode global joint relations and local.

bone connectivity; Zhai et al. [44] present HGFreNet, combining hop-based graph attention blocks with Transformer encoders and enforcing temporal consistency via frequency-domain loss for smoother trajectories. Frequency-domain techniques, such as graph wavelets and scattering, mitigate GCN low-pass effects and enhance long-sequence efficiency by preserving high-frequency components [21, 22]. While these hybrid and frequency-domain methods advance performance on multiple fronts, existing works often face challenges in simultaneously achieving strong structural awareness and effective recovery of high-frequency geometric details within a lightweight framework [45–48].

Based on the above analysis, the proposed SBAHGNet adopts a dual-branch design for integration: the skeleton-biased attention branch explicitly embeds topology into self-attention to enhance structural awareness; The proposed unified and lightweight framework directly addresses the two primary limitations commonly observed in prior approaches (particularly hybrid methods): insufficient explicit structural awareness of the human skeleton and inadequate preservation of high-frequency geometric details. This design confers substantial advantages over existing hybrid paradigms in terms of model simplicity, computational efficiency, and performance on fine-grained motion capture.

3 Method

3.1 Overall architecture

As shown in Fig. 1, the overall architecture of the proposed.

SBAHGNet is as follows. The model input is a 2D keypoint sequence $x \in R^{B\times T\times J\times 3}$, and the model output (prediction) is $\hat{P} \in R^{B\times T\times J\times 3}$, where B is the batch size, T is the number of frames, and J is the number of joints. The last dimension 3 denotes the 2D coordinates plus a confidence score. The data processing pipeline is: first, the input joint features x are linearly projected along the last dimension to D channels, yielding features $X \in R^{B\times T\times J\times D}$. A spatial positional encoding $P^s \in R^{1\times J\times D}$is then added to X (broadcasted over the B and T dimensions). The resultant representation (denoted as $F^{(0)} \in R^{B\times T\times J\times D}$) is subsequently input to N cascaded SBAHGBlock modules, thereby progressively extracting hierarchical spatio-temporal features, producing $F^{(N)} \in R^{B\times T\times J\times D}$. Finally, the joint features are mapped to a higher dimensional space via linear layers and a regression head is used to produce the predicted 3D keypoint sequence $\hat{P}$. The present model is trained by employing both a position loss function (L_{3D}) and a velocity loss function ($L_{\Delta P}$). The corresponding formulas are given as follows:

$$\begin{aligned} L_{3D} &= \sum_{\mathrm{t}=1}^{T}\sum_{j=1}^{J} ||\hat{P}_{\mathrm{t},j} - P_{t,j}|| \\ L_{\Delta P} &= \sum_{\mathrm{t}=2}^{T}\sum_{j=1}^{J} ||\Delta\hat{P}_{\mathrm{t},j} - \Delta P_{t,j}|| \end{aligned} \tag{1}$$

Where $\Delta\hat{P}_{t} = \hat{P}_{t} - \hat{P}_{t-1}$ denotes the inter-frame differences of the predicted 3D pose sequence output by the model, and $\Delta P_t = P_t - P_{t-1}$ represents the inter-frame differences of the corresponding ground-truth 3D pose sequence. The overall loss function of the model is defined as follows:

$$L = L_{3D} + \lambda_{\Delta P} L_{\Delta P} \tag{2}$$

where $\lambda_{\Delta_P}$is a hyperparameter that balances positional accuracy and motion smoothness.

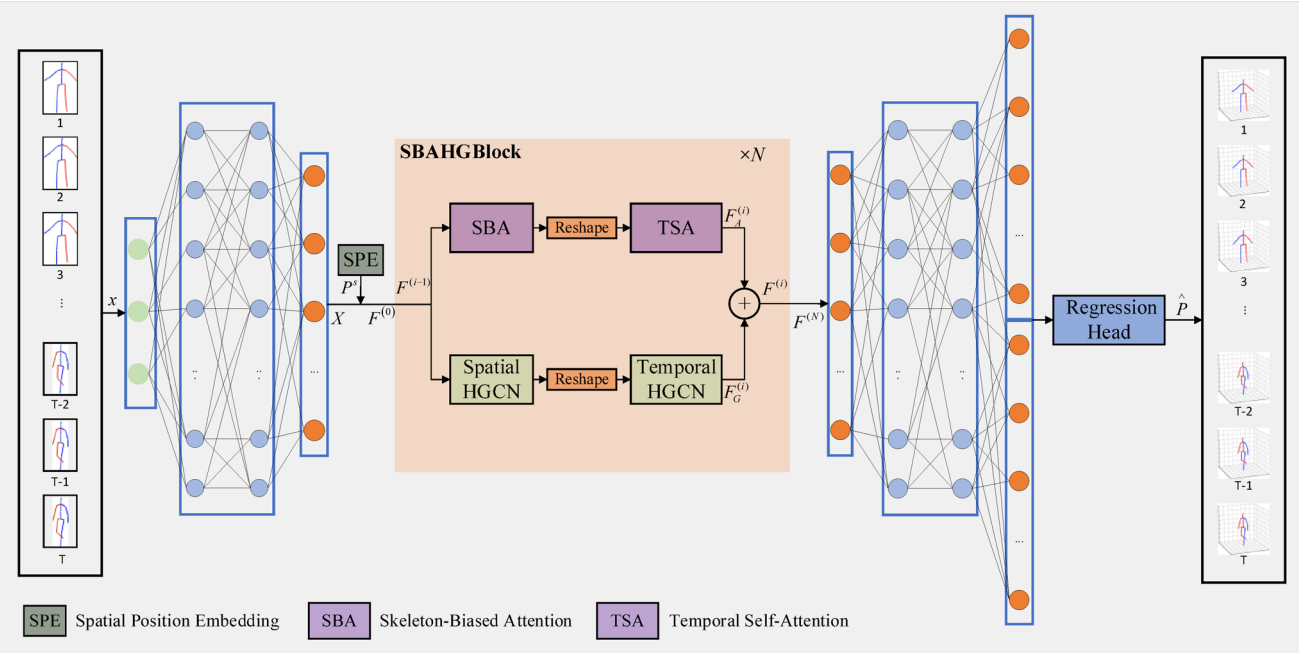


Fig. 1 The overall architecture of SBAHGNet

3.2 SBAHGBlock

The proposed module is composed of three primary components: a GCN branch, an attention branch, and an adaptive fusion module. Within the GCN branch, a temporal High-Frequency Enhanced Graph Convolutional Network (HGCN) module and a spatial HGCN module are employed, both following the architecture depicted in Fig. 2. The attention branch consists of a Skeletal-Biased Attention (SBA) module and a Temporal Self-Attention (TSA) module. Both branches adopt a cascaded architecture, where spatial features are extracted prior to temporal features. The GCN branch employs a spatio-temporal HGCN module to extract and fuse spatio-temporal adjacency relationships, thereby enhancing the spatio-temporal representation of 3D poses. The attention branch utilizes SBA and TSA to capture global information and effectively model long-range dependencies in human motion, wherein the skeleton bias introduced in SBA facilitates a better understanding of spatial dependencies among joints. Finally, features from the two branches are integrated via the adaptive fusion module, yielding a fused representation that balances different information foci.

3.2.1 Spatio-temporal HGCN module

The spatial HGCN and temporal HGCN are respectively used to extract spatial and temporal information from human motion, thereby capturing local connectivity relationships between joints as well as temporal dependencies during the motion process. The difference between the spatial module and the temporal module lies in the input to the spatio-temporal module and the adjacency matrix. The output $F_{GS}^{(i)}$ of the spatial HGCN and the output $F_{G}^{(i)}$ of the temporal HGCN are formulated as shown in Eqs. (3) and (4), respectively.

$$\begin{aligned} G_S &= \mathrm{DP}(\mathrm{Reshape}(W(\hat{A}\,\mathbf{V}(F_R^{(i-1)}))) + \mathbf{U}(F_R^{(i-1)})) \\ GCN(F_{GS}^{(i)}) &= \mathrm{Reshape}(\mathrm{LBR}(G_S + \mathrm{MHSA}(G_S))) \end{aligned} \tag{3}$$

$$\begin{aligned} G_T &= \mathrm{DP}(\mathrm{Reshape}(W(\hat{A}\,\mathbf{V}(F_{RGS}^{(i)}))) + \mathbf{U}(F_{RGS}^{(i)})) \\ GCN(F_{G}^{(i)}) &= \mathrm{Reshape}(\mathrm{LBR}(G_T + \mathrm{MHSA}(G_T))) \end{aligned} \tag{4}$$

Where $F_R^{(\mathrm{i}-1)} \in R^{(BT)\times J\times D}$ represents the output of the previous.

SBAHGBlock after sequence rearrangement. For the spatial module, the input is $F_R^{(\mathrm{i}-1)}$ and the output $F_{GS}^{(i)} \in R^{B\times T\times J\times D}$ denotes the output of the current spatial module, with the adjacency matrix constructed based on human body topology. For the temporal module, the input is $F_{RGS}^{(\mathrm{i})} \in R^{(BJ)\times T\times D}$ (i.e., the rearranged $F_{GS}^{(i)}$) and the output $F_G^{(i)} \in R^{B\times T\times J\times D}$ represents the final output of the current branch, with the adjacency matrix constructed based on inter-frame joint similarity. $\mathbf{V}$ and $\mathbf{U}$ are two trainable weight matrices, $\hat{A} = A + I_N$ represents the adjacency matrix with self-connections added, I_N stands for the identity matrix, $W(\cdot)$ denotes the processing by MSHFE,

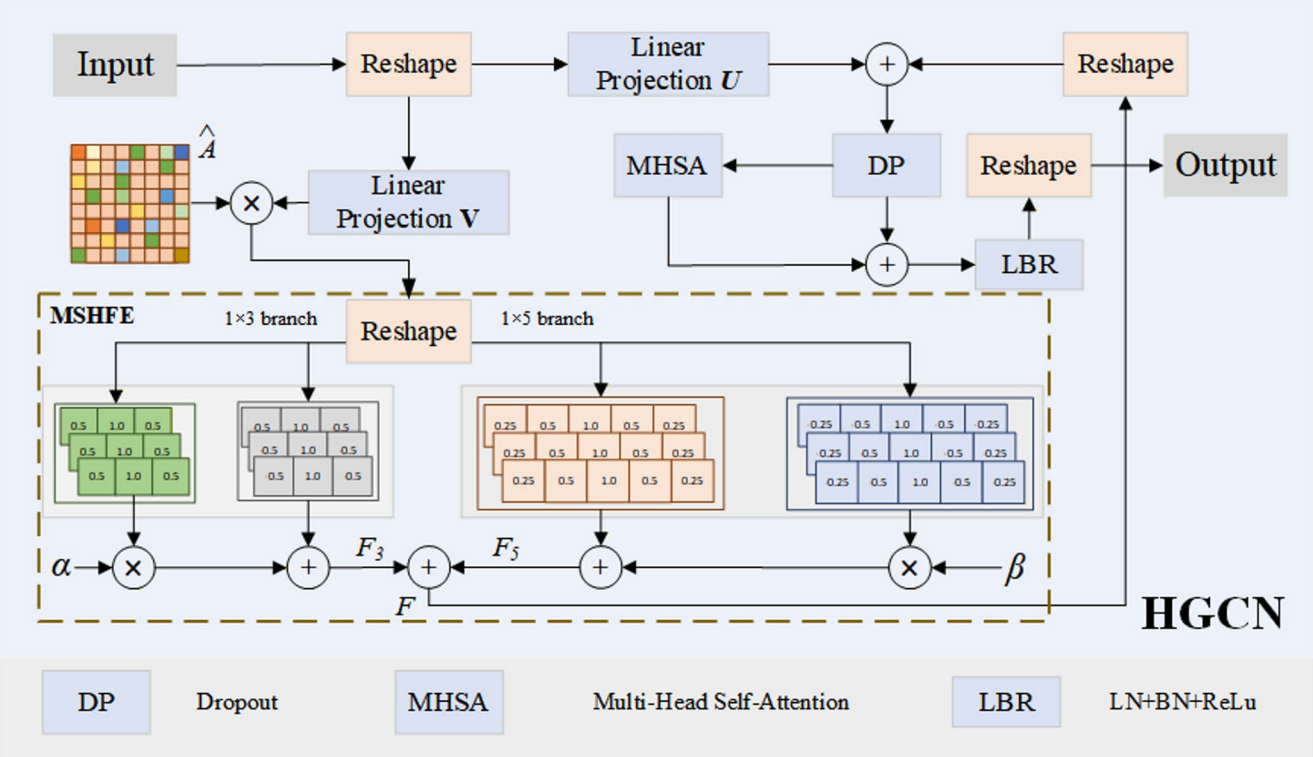


Fig. 2 The overall architecture of HGCN, where the brown dashed portion illustrates the structure of the MSHFE module

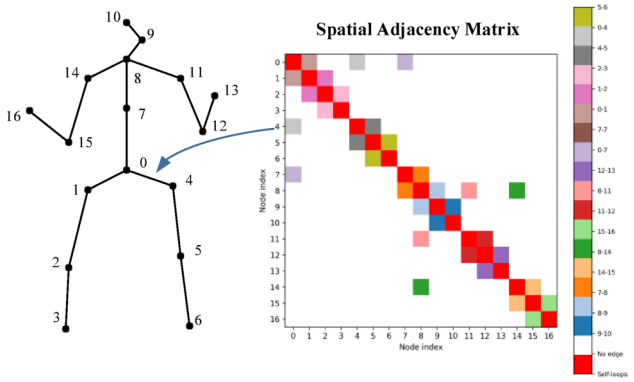


Fig. 3 Spatial adjacency matrix based on human skeletal topology

and LBR refers to the sequential operations of layer normalization, batch normalization, and ReLU activation.

For the construction of the adjacency matrix A, in the spatial module, the adjacency matrix is predefined based on the topological relationships of human joints, with the specific connectivity illustrated in Fig. 3. In the temporal module, we employ a dynamic adjacency matrix: first, the input temporal sequence features are L2-normalized; then, cosine similarity between all pairs of time steps is computed via inner product (i.e., cosine similarity is calculated for the same joint across different frames), forming a similarity matrix; subsequently, for each row, the top-K largest values are selected as the threshold, and edges with similarity greater than or equal to this threshold are set to 1, yielding a sparse binary adjacency matrix; finally, symmetric normalization is applied for use in graph convolution. By combining the dynamic adjacency matrix with the graph convolutional network, the model can fully account for temporal relationships between different time steps during learning, thereby better capturing long-range temporal dependencies. The advantage of this approach lies in its ability to adaptively construct the adjacency matrix based on the input data at each moment, providing greater flexibility and adaptability across diverse motion scenarios. Figure 4 illustrates an example with 3 temporal frames, where a K-nearest neighbors (KNN) strategy is adopted with K = 1, converting the similarity matrix into a binary adjacency matrix

Traditional graph convolutional networks exhibit an inherent smoothing effect during feature aggregation, which leads to the loss of high-frequency details in temporal sequences, such as motion boundaries and joint accelerations [18–20]. To mitigate this issue, we propose MSHFE, which is embedded in the aggregation stage of graph convolution. This module extracts local dynamic features through multi-scale convolution and enhances high-frequency information using an adaptive mechanism, thereby improving the model's expressive capability for complex motion sequences. The structure of MSHFE is illustrated in Fig. 2.

MSHFE employs a dual-branch parallel architecture, utilizing 1×3 and 1×5 depthwise separable convolutions for

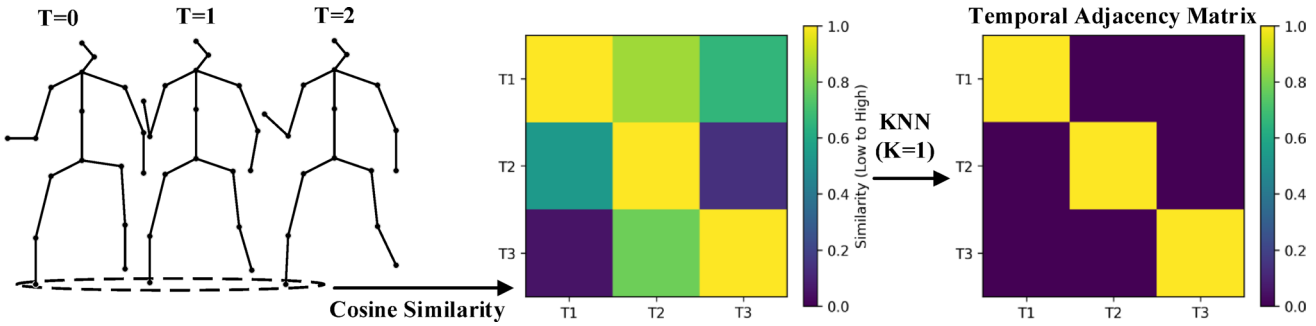


Fig. 4 Schematic illustration of the Temporal Adjacency Matrix constructed using the K-Nearest Neighbors (K-NN) algorithm. Connected edges are determined by considering the highest similarity of each joint across the entire temporal sequence (e.g., the right ankle in the figure)

efficient multi-scale feature decomposition. The convolution stride is 1 for both, with padding of 1 and 2, respectively, to ensure that the output sequence length matches the input exactly and avoids boundary information loss. Each branch consists of low-frequency and high-frequency sub-paths:

(1) Low-frequency sub-path

A predefined smoothing kernel (1 × 3: [0.5, 1.0, 0.5]; 1 × 5: [0.25, 0.5, 1.0, 0.5, 0.25]) is applied to perform local weighted averaging on the input sequence, preserving the structural backbone information.

(2) High-frequency sub-path

Corresponding difference kernels (1 × 3: [– 0.5, 1.0, – 0.5]; 1 × 5: [– 0.25, – 0.5, 1.0, – 0.5, – 0.25]) are used to capture local rates of change. The high-frequency output is multiplied by learnable high-frequency gain factors α and β, which are initialized in logarithmic form to ensure positivity and numerical stability. These gains are dynamically adjusted during training to achieve adaptive amplification of abrupt motions.

The low-frequency and high-frequency outputs at each scale are concatenated along the channel dimension, followed by 1 × 1 convolution, batch normalization, and ReLU activation for non-linear fusion within the same scale, producing refined representations F_3 and F_5. Subsequently, the module introduces a lightweight cross-scale attention mechanism: first, F_3 and F_5 are averaged and subjected to global average pooling to obtain a global context descriptor; then, weights are generated through two layers of 1 × 1 convolutions ($C \rightarrow C/2 \rightarrow 2$) followed by Softmax normalization. The final output is:

$$F = \alpha_3 \cdot F_3 + \alpha_5 \cdot F_5 \tag{5}$$

3.2.2 Spatio-temporal attention module

The attention branch is responsible for extracting spatio-temporal features from the input, primarily achieved through a two-stage multi-head self-attention mechanism that processes spatial information first and temporal information second. Specifically, this branch first models spatial relationships between joints within each frame, followed by capturing long-range dependencies along the temporal dimension for the motion trajectory of each joint. This sequential design facilitates the separate handling of intra-frame structural information and inter-frame dynamic information. In this paper, a lightweight multi-scale skeleton bias is introduced in the spatial attention stage of this branch to provide additional skeletal topological priors for the attention weights; the temporal attention stage retains the standard implementation without any additional operations. The formulation of the spatial SBA is as follows:

$$\begin{aligned} SBA(Q_s, K_s, V_s) &= Concat(head_1, \ldots, head_h)W_S^O \\ head_i &= softmax(\frac{Q_s^{(i)}(K_s^{(i)})^T}{\sqrt{d_k}} + B)V_s^{(i)} \end{aligned} \tag{6}$$

where W_S^O is the output projection matrix, h is the number of parallel attention heads, $B \in R^{B\times T\times J\times J}$ denotes the multi-scale skeleton bias (its computation will be elaborated in the subsequent section), and $d_k = D/h$ represents the channel dimension per attention head. To compute the query matrix $Q_s^{(i)}$, key matrix $K_s^{(i)}$, and value matrix $V_s^{(i)}$, we have

$$Q_s^{(i)} = F_S W_s^{Q,i}, K_s^{(i)} = F_S W_s^{K,i}, V_s^{(i)} = F_S W_s^{V,i} \tag{7}$$

where $F_S \in R^{BT\times J\times D}$ is the reshaped version of the output $F^{(i-1)}$ from the previous layer, and $W_s^{Q,i}$,$W_s^{K,i}$, $W_s^{V,i}$ are the projection matrices.

The multi-scale skeleton bias B interacts with the input features $F^{(i-1)}$ through the association matrix $H_s \in R^{E_s\times J}$ to capture higher-order relationships between different

joints. This bias is based on a normalized association matrix (allowing fine-tuning during training) and learnable scale weights w_s, injecting skeletal topological priors into the attention mechanism in the form of a dynamic outer product, thereby providing additional positive offsets for structurally related joint pairs. The bias is shared across all attention heads and broadcast along the head dimension before being added to the original dot-product scores. This process does not alter the standard attention computation flow and serves merely as a lightweight additive term for biasing. The specific computation is divided into three steps, with the formulations as follows:

1. Node-to-Hyperedge Aggregation

For the s-th scale, the normalized incidence matrix H_s is utilized to average the joint features within the same hyperedge, generating the hyperedge feature $I_e \in R^{B\times T\times E_s\times D}$:

$$I_e = H_s \cdot F^{(i-1)} \tag{8}$$

where $s \in \{1, 2\}$ denotes the scale index,E_s is the number of hyperedges at this scale, and J=17 is the number of joints. Each row of the matrix corresponds to a hyperedge e; if joint j belongs to hyperedge e, then $H_s(e, j) = 1/|e|$, otherwise 0. The matrix is initialized and treated as a learnable parameter, allowing fine-tuning during training. The specific initializations of the incidence matrices for the two scales are shown in Fig. 5.

(2) Hyperedge-to-node broadcasting

The aggregated hyperedge features are broadcast back to the individual joints, yielding smoothed node features $B_e \in R^{B\times T\times J\times D}$:

$$B_e = H_s^T \cdot F^{(i-1)} \tag{9}$$

where B_e represents the bias information broadcast from hyperedge e, capturing similarity relationships between joints.

(3) Outer product and scale fusion

The outer product of the smoothed features B_e and the original features $F^{(i-1)}$ is computed, ensuring that joints belonging to the same hyperedge share the same broadcast component, thereby naturally receiving additional positive contributions during similarity computation. Subsequently, the results from the two scales are fused using learnable weights w_s (normalized via Softmax) to obtain the final skeleton bias $B \in R^{B\times T\times J\times J}$:

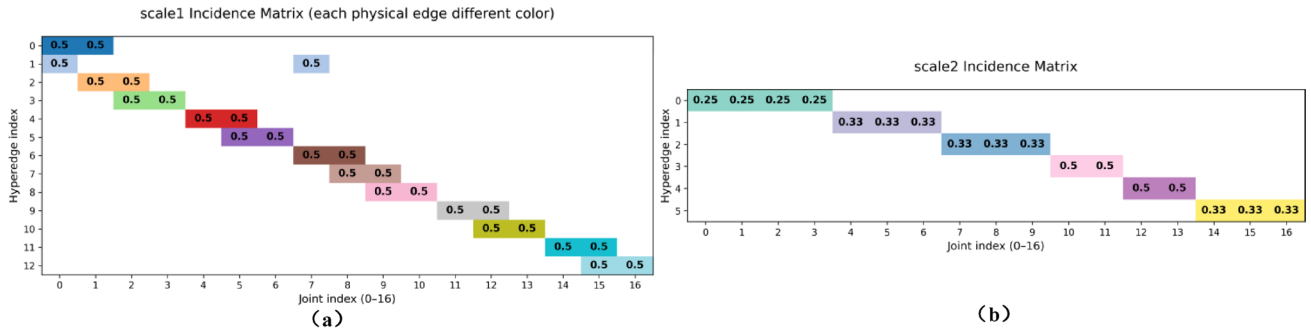
$$B = \sum_{i=1}^{s} w_s(B_e(F^{(i-1)})^T) \tag{10}$$

The complete formulation integrating the above three steps is given as follows:

$$B = \sum_{i=1}^{s} w_s((H_s^T(H_s F^{(i-1)}))(F^{(i-1)})^T) \tag{11}$$

Subsequently, the output of the spatial SBA is reshaped into $F_T \in R^{BJ\times T\times D}$ to serve as the input to the TSA, which is used to extract temporal features of the joints. The computation formula for the temporal TSA is similar to that of the spatial SBA without the bias B, and is given as follows:

$$\begin{aligned} TSA(Q_T, K_T, V_T) &= Concat(head_1, ..., head_h)W_T^O \\ head_i &= softmax(\frac{Q_t^{(i)}(K_t^{(i)})^T}{\sqrt{d_k}})V_t^{(i)} \end{aligned} \tag{12}$$



13×17 6×17

Fig. 5 Initialization of the normalized incidence matrices, where different colors represent distinct hyperedges. **a** Fine-scale normalized incidence matrix (13 × 17); **b** Coarse-scale normalized incidence matrix (6 × 17)

where Q_T, K_T, and V_T are computed in a similar manner to Eq. (7) (i.e., obtained through corresponding linear projections).

3.2.3 Adaptive fusion

To fully leverage the complementary modeling capabilities of the attention branch and the graph convolution branch, we incorporate a lightweight adaptive fusion module in each SBAHGBlock. Unlike simple averaging or concatenation operations, this module dynamically assigns weights from the two branches for each joint and each frame. Specifically, after both branches complete their respective spatial and temporal processing, the output $F_A^{(i)}$ from the attention branch and the output $F_G^{(i)}$ from the graph convolution branch are concatenated along the channel dimension to obtain $F_{\text{cat}} \in R^{B\times T\times J\times 2D}$. Subsequently, a linear layer with only $2C+2$ parameters is applied to project and generate the weights. The final fused output $F^{(i)}$ is computed as follows:

$$\begin{aligned} \alpha &= \text{Softmax}(W_f F_{cat} + b_f) \in R^{B\times T\times J\times 2} \\ F^{(i)} &= \alpha[\cdots, 0:1] \odot F_A^{(i)} + \alpha[\cdots, 0:2] \odot F_G^{(i)} \end{aligned} \tag{13}$$

where W_f and b_f are the weight matrix and bias vector of the projection linear layer, respectively, $\odot$denotes element-wise multiplication, and the weight α is automatically broadcast along the channel dimension to match the feature dimension D.

4 Experiments

4.1 Datasets and evaluation metrics

To evaluate the performance of the proposed model, we conduct experiments on two widely used benchmark datasets for 3D human pose estimation: Human3.6M [49] and MPI-INF-3DHP [50].

Human3.6 M is currently the most widely used benchmark dataset for indoor human pose estimation, comprising approximately 3.6 million frames captured from 11 subjects performing 15 categories of daily activities. Consistent with previous studies [16, 39], to ensure comparability with existing works, we adopt the standard data split: training on subjects 1, 5, 6, 7, and 8, and testing on subjects 9 and 11.Evaluation is conducted using two standard protocols: Protocol #1 (MPJPE) measures the mean per-joint position error (in millimeters) after aligning the root joints of the predicted and ground-truth poses; Protocol #2 (P-MPJPE) computes the error after rigid Procrustes alignment between the predicted pose and the ground truth.

MPI-INF-3DHP includes both indoor and outdoor scenes, with its test set covering three types of environments: studio with green screen, studio without green screen, and outdoor. Following the practices of previous works, we report the mean per-joint position error (MPJPE), the percentage of correct keypoints (PCK) with a threshold of 150 mm, and the corresponding area under the curve (AUC). These metrics collectively provide a comprehensive reflection of the method's performance in terms of spatial accuracy and robustness in keypoint detection.

4.2 Implementation details

The proposed SBAHGNet model is implemented using PyTorch and trained on a single NVIDIA RTX 3080 Ti GPU. Horizontal flipping is applied as data augmentation during both training and testing phases, following [51, 52]. During training, the batch size is set to 2. The AdamW [53] optimizer is employed for network parameter optimization, with training conducted for 120 epochs and a weight decay of 0.01. The initial learning rate is set to 5e-4, with an exponential decay schedule applied using a decay factor of 0.99.

For experiments on Human3.6 M, 2D pose inputs are obtained from either the Stacked Hourglass detector [54] or the ground-truth 2D poses provided by the dataset. For MPI-INF-3DHP, ground-truth 2D poses from the dataset are used as input. Other key hyperparameters include a feature dimension D of 128, 8 attention heads ($h=8$) in the attention branch, and 2 nearest neighbors in the temporal adjacency matrix.

4.3 Comparison experiment

Results on Human3.6M. We compare the proposed SBAHGNet with other methods on the Human3.6 M dataset (as shown in Table 1). To ensure a fair comparison, only results from models without pre-training on additional data are included. SBAHGNet achieves an MPJPE of 37.2 mm with estimated 2D pose inputs and 12.4 mm with ground-truth 2D pose inputs. Notably, compared to MotionBERT [56], our method utilizes only 43% of its parameter count and 51% of its computational resources, while improving accuracy by 2.0 mm and 5.4 mm, respectively. Furthermore, compared to another state-of-the-art model, TCPFormer [59], our model employs only 52% of its parameters yet achieves accuracy improvements of 0.7 mm and 3.1 mm, respectively.

Results on MPI-INF-3DHP. We further evaluate the generalization capability of the model on the more challenging MPI-INF-3DHP dataset. Due to the shorter video sequences

Table 1 Quantitative comparisons on Human3.6 M

Method	T	CE	Param	MACs	P1↓/P2↓	P1†↓
MHFormer [39] CVPR'22	351	√	30.9 M	7.0G	43.0/34.3	30.5
P-STMO [55] ECCV'22	243	√	6.2 M	0.7G	42.8/34.4	29.3
STCFormer [40] CVPR'23	243	×	4.7 M	19.6G	41.0/32.0	21.3
STCFormer-L [40] CVPR'23	243	×	19.9 M	78.2G	40.5/31.8	-
PoseFormerV2 [28] CVPR'23	243	√	14.4 M	4.8G	45.2/35.6	-
GLA-GCN [13] ICCV'23	243	√	1.3 M	1.5G	44.4/34.8	21.0
MotionBERT [56] ICCV'23	243	×	42.3 M	174.8G	39.2/32.9	17.8
HDFormer [57] IJCAI'23	96	×	3.7 M	0.6G	42.6/33.1	21.6
HSTFormer [58] arXiv'23	81	×	22.7 M	1.0G	42.7/33.7	27.8
MotionAGFormer-L [14] WACV'24	243	×	19.0 M	78.3G	38.4/32.5	17.3
TCPFormer [59] AAAI'25	243	×	35.1 M	109.2G	*37.9/31.7*	*15.5*
SBAHGNet	243	×	18.3 M	88.9G	**37.2/31.6**	**12.4**

T: Number of input frames. CE: Estimating center frame only. P1: MPJPE error (mm). P2: P-MPJPE error (mm). P1†: P1 error on 2D ground truth. (*) denotes using HRNet [62] for 2D pose estimation. The best and second-best scores are in bold and underlined, respectively. For detailed per-action results, please refer to Table 3 as shown

Table 2 Quantitative comparisons on MPI-INF-3DHP

Method	T	PCK↑	AUC↑	P1↓
MHFormer [39] CVPR'22	9	93.8	63.3	58.0
P-STMO [55] ECCV'22	81	97.9	75.8	32.2
STCFormer [40] CVPR'23	81	*98.7*	83.9	23.1
PoseFormerV2 [28] CVPR'23	81	97.9	78.8	27.8
GLA-GCN [13] ICCV'23	81	98.5	79.1	27.7
MotionBERT [56] ICCV'23	–	–	-	-
HDFormer [57] IJCAI'23	96	98.7	72.9	37.2
HSTFormer [58] arXiv'23	81	97.3	71.5	41.4
MotionAGFormer-L [14] WACV'24	81	98.2	85.3	16.2
TCPFormer [59] AAAI'25	81	**99.0**	*87.7*	*15.0*
SBAHGNet	81	**99.0**	**88.2**	**13.8**

T: Number of input frames. The best and second-best scores are in bold and underlined, respectively. (ties are marked accordingly)

in this dataset, we adjust the input frame length to 81 frames. As shown in Table 2, our model delivers superior performance, attaining a PCK of 99.0%, an AUC of 88.2%, and an MPJPE of 13.8 mm, surpassing the current best model TCPFormer (with a 0.5% improvement in AUC and a 1.2 mm reduction in MPJPE).

4.4 Ablation experiments

We conducted a series of ablation experiments on the Human3.6 M dataset to systematically analyze the effectiveness of different architectural design choices in SBAHGNet.

Model Depth and Width. We analyzed the impact of the number of stacked SBAHGBlock modules N and the input feature dimension D on the model's performance in terms of P1 and P2 metrics, with results shown in Table 4. Overall, as the number of modules increases and the feature dimension grows, the model performance exhibits a consistent improvement trend. Additionally, increasing the feature dimension D significantly escalates the model's parameter count and computational overhead. Notably, when the number of modules is 18 and the input feature dimension is 64, the model achieves comparable performance to TCPFormer on the P1 metric while utilizing only approximately 13% of its parameters and 24% of its computational resources.

Impact of MSHFE and Multi-Scale Skeleton Bias. To systematically investigate the effects of our proposed MSHFE module and Multi-Scale Skeleton Bias on model performance, we conducted two sets of controlled ablation experiments on the Human3.6 M dataset: overall performance evaluation for global contribution validation, and limb joint-level MPJPE analysis for fine-grained mechanism interpretation. A unified baseline model is adopted for all experiments, which refers to the SBAHGNet backbone with both the MSHFE module and multi-scale skeleton bias removed. All experiments strictly follow the single-variable control principle: only the target module is adjusted, while all other training and inference configurations remain exactly the same to ensure fair and reliable comparisons. We first validated the overall contribution of the two modules via global performance ablation tests, with the results summarized in.

Table 5. When both augmentation modules are removed, the model suffers the most severe performance degradation, with P1 and P2 errors increasing by 1.2 mm and 0.8 mm, respectively. Incorporating only the MSHFE module yields a certain degree of performance improvement, while introducing only the Multi-Scale Skeleton Bias also achieves corresponding performance optimization. This indicates that both modules bring positive gains to the model's feature representation capability, and their combination delivers a synergistic effect. To further dissect the differentiated working mechanism and functional complementarity of the two modules at the joint level, we conducted fine-grained ablation experiments on the MPJPE of each limb joint, with

Table 3 Quantitative comparisons of 3D human pose estimation per action on Human3.6 M

MPJPE	T	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
*MHFormer [39] CVPR'22	351	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
P-STMO [55] ECCV'22	243	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
STCFormer [40] CVPR'23	243	39.6	41.6	37.4	38.8	43.1	51.1	39.1	39.7	51.4	57.4	41.8	38.5	40.7	27.1	28.6	41.0
STCFormer-L [40] CVPR'23	243	38.4	41.2	*36.8*	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	*26.2*	27.7	40.5
UPS [60] CVPR'23	243	37.5	39.2	36.9	40.6	**39.3**	*46.8*	39.0	41.7	50.6	63.5	40.4	37.8	44.2	26.7	29.1	40.8
GLA-GCN [13] ICCV'23	243	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
†MotionBERT [56] ICCV'23	243	36.6	39.3	37.8	33.5	41.4	49.9	37.0	35.5	50.4	56.5	41.4	38.2	37.3	*26.2*	26.9	39.2
HDFormer [57] IJCAI'23	96	38.1	43.1	39.3	39.4	44.3	49.1	41.3	40.8	53.1	62.1	43.3	41.8	43.1	31.0	29.7	42.6
HSTFormer [58] arXiv'23	81	39.5	42.0	39.9	40.8	44.4	50.9	40.9	41.3	54.7	58.8	43.6	40.7	43.4	30.1	30.4	42.7
MotionAGFormer-L [14] WACV'24	243	36.8	38.5	**35.9**	33.0	41.1	48.6	38.0	*34.8*	49.0	*51.4*	40.3	37.4	*36.3*	27.2	27.2	38.4
TCPFormer [59] AAAI'25	243	*36.4*	*37.7*	**35.9**	*32.6*	40.6	47.3	*36.7*	*34.8*	**47.7**	52.3	*39.9*	*36.8*	36.6	26.6	*26.8*	*37.9*
SBAHGNet	243	**35.2**	**37.6**	37.9	**31.7**	*39.7*	**45.7**	**36.3**	**33.8**	*48.7*	**50.0**	**39.5**	**35.6**	**35.7**	**25.3**	**25.9**	**37.2**
P-MPJPE	T	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
*MHFormer [39] CVPR'22	351	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
P-STMO [55] ECCV'22	243	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
STCFormer [40] CVPR'23	243	*29.5*	33.2	**30.6**	31.0	33.0	38.0	30.4	*29.4*	41.8	45.2	33.6	29.5	31.6	*21.3*	22.6	32.0
STCFormer-L [40] CVPR'23	243	**29.3**	33.0	*30.7*	30.6	*32.7*	38.2	**29.7**	**28.8**	42.2	*45.0*	*33.3*	*29.4*	31.5	**20.9**	*22.3*	31.8
UPS [60] CVPR'23	243	30.3	*32.2*	30.8	33.1	**31.1**	**35.2**	30.3	32.1	**39.4**	49.6	**32.9**	**29.2**	33.9	21.6	24.5	32.5
GLA-GCN [13] ICCV'23	243	32.4	35.3	32.6	34.2	35.0	42.1	32.1	31.9	45.5	49.5	36.1	32.4	35.6	23.5	24.7	34.8
†MotionBERT [56] ICCV'23	243	30.8	32.8	32.4	28.7	34.3	38.9	30.1	30.0	42.5	49.7	36.0	30.8	31.7	22.0	23.0	32.9
HDFormer [57] IJCAI'23	96	29.6	33.8	31.7	31.3	33.7	37.7	30.6	31.0	41.4	47.6	35.0	30.9	33.7	25.3	23.6	33.1
HSTFormer [58] arXiv'23	81	31.1	33.7	33.0	33.2	33.6	38.8	31.9	31.5	43.7	46.3	35.7	31.5	33.1	24.2	24.5	33.7
MotionAGFormer-L [14] WACV'24	243	31.0	32.6	31.0	27.9	34.0	38.7	31.5	31.5	41.4	45.4	34.8	30.8	*31.3*	22.8	23.2	32.5
TCPFormer [59] AAAI'25	243	30.1	**31.6**	31.4	*27.3*	33.5	37.6	29.4	29.6	*41.1*	45.9	34.4	29.6	**30.6**	21.7	*22.3*	*31.7*
SBAHGNet	243	29.7	**31.6**	32.6	**27.0**	33.3	*36.7*	*29.9*	*29.4*	41.4	**43.8**	34.5	29.7	**30.6**	*21.3*	**22.0**	**31.6**

(Top) MPJPE (mm) using detected 2D pose sequence. (Bottom) P-MPJPE (mm) using detected 2D pose sequence. (*) denotes using HRNet [62] for 2D pose estimation. (†) denotes manually evaluated using their provided evaluation code. The best results are highlighted in bold, and the second-best results are underlined (ties are marked accordingly)

Table 4 The P1 error comparison by varying the number of SBAHGBlock modules and the number of channels. D denotes the number of channels in each SBAHGBlock module. T is kept at 243 in all experiments

N	D	Param	MACs	P1↓
9	64	2.4 M	13.2G	39.1
12	64	3.1 M	17.5G	38.7
18	64	4.7 M	26.2G	37.9
9	128	9.2 M	44.6G	38.1
12	128	12.3 M	59.4G	38.0
16	128	16.3 M	79.1G	37.5
17	128	17.3 M	84.0G	37.5
18	128	18.3 M	88.9G	37.2

Table 5 Ablation study on the impact of MSHFE and multi-scale skeleton bias

MSHFE	Multi-scale skeleton bias (B)	P1↓/P2↓
×	×	38.4/32.4
×	√	37.9/32.1
√	×	38.1/32.3
√	√	37.2/31.6

detailed results presented in Table 6. The MSHFE module achieves a noticeable reduction in estimation errors for high-frequency motion joints (i.e., knee and elbow joints), with a maximum MPJPE decrease of 3.7 mm for the right knee. This result aligns with our design motivation, verifying that the module has a certain alleviation effect on the high-frequency geometric detail loss caused by the inherent low-pass behavior of GCN. In comparison, the multi-scale skeleton bias module delivers more prominent performance gains for upper limb joints, forming a clear functional complementarity with the MSHFE module. Benefiting from this complementary design, the full SBAHGNet model achieves the optimal estimation accuracy across all limb joints, which is consistent with the conclusion of the global performance tests and fully validates the rationality of our module design.

Effect of Temporal and Spatial Positional Embeddings. To analyze the impact of positional embeddings on model performance, we conducted corresponding ablation experiments, with results shown in Table 7. The experiments reveal that adding temporal positional embeddings on top of spatial positional embeddings actually increases the P1 error. We attribute this phenomenon to the permutation-equivariant property of the graph convolution branch, which

Table 7 Accuracy error comparison of temporal position embedding and spatial position embedding

Spatial embedding	Temporal embedding	P1
×	×	37.9
√	×	37.2
×	√	37.7
√	√	38.2

Table 8 Ablation study on GCN branch and Attention branch

GCN	Attention	P1↓
√	×	38.6
×	√	38.5
√	√	37.2

enables the network to naturally preserve frame-level temporal order during modeling, thereby reducing the necessity of explicit temporal positional embeddings.

Ablation on Branch Contributions. We further conducted ablation experiments on the two branches, with results presented in Table 8. Notably, using the graph convolution branch or the attention branch alone yields nearly equivalent performance (38.6 mm vs. 38.5 mm), indicating comparable standalone modeling capabilities for each. Through adaptive fusion of the dual branches, the error is further reduced to 37.2 mm, representing a 1.3 mm improvement over the stronger single-branch baseline.

4.5 Qualitative comparison and analysis

Figure 6 presents the visualization results for the Walking, Smoking, and Sitting actions of subject S9 on the Human3.6 M test set. It can be observed that, compared to existing state-of-the-art methods-including TCPFormer, MotionAGFormer, and PoseFormerV2, our approach generates more accurate and consistent 3D human pose predictions. The black circles highlight regions where other methods exhibit significant prediction errors at local joints, further underscoring the superiority of the proposed model in fine-grained pose modeling. To further evaluate the generalization performance of the proposed method in real-world scenarios, we applied the pre-trained model to in-the-wild outdoor videos. Specifically, we first employed YOLOv3 [61] for human detection in the videos, followed by HRNet [62] to extract 2D keypoints. Subsequently, the SBAHGNet

Table 6 Ablation study on MPJPE of limb joints on the Human3.6 M dataset

MPJPE	Left hip	Left knee	Left foot	Right hip	Right knee	Right foot	Right shoulder	Right elbow	Right hand	Left shoulder	Left elbow	Left hand
Baseline	*20.5*	36.4	*52.4*	21.1	37.9	57.3	39.8	47.5	58.1	41.7	53.6	62.8
MSHFE only	20.9	**33.9**	55.2	*20.5*	*34.2*	57.9	*37.8*	*46.7*	58.0	**39.1**	**51.2**	**60.8**
Multi-Scale Skeleton Bias only	21.0	34.3	53.3	21.4	34.4	*56.6*	38.4	**45.9**	*57.7*	40.2	*51.3*	*62.4*
SBAHGNet	**20.1**	*33.6*	**52.1**	**20.0**	**33.6**	**55.5**	**36.4**	*46.7*	**57.2**	*40.1*	53.1	62.6

The best results are highlighted in bold, and the second-best results are underlined

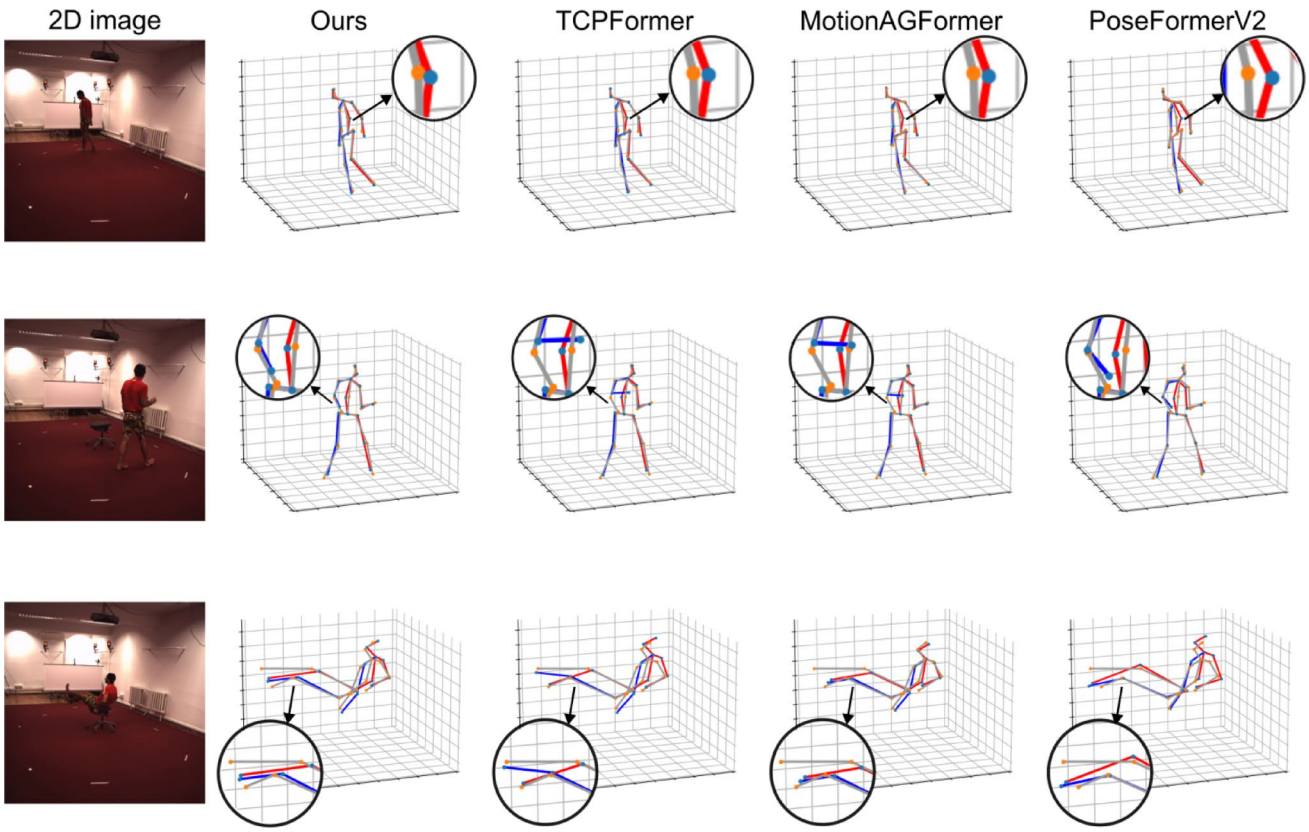


Fig. 6 Qualitative Comparison Results on the Human3.6 M Dataset: The gray skeleton represents the ground-truth 3D pose, with the actual keypoints shown in orange. The blue color denotes the torso and the left side of the body, while the red color indicates the estimated right side. The predicted keypoints are depicted in blue

model, pre-trained on the Human3.6 M dataset, was utilized to recover 3D human poses from randomly selected challenging video sequences involving ice skating, dancing, and Tai Chi actions. As shown in Fig. 7, SBAHGNet maintains stable and accurate predictions even in complex and diverse real-world scenarios, demonstrating its strong generalization capability and potential for practical application. Furthermore, we further evaluate the performance of our model under occluded scenarios. As shown in Fig. 8, SBAHGNet is still able to generate reasonable 3D pose estimates under occluded conditions on the Human3.6 M dataset, demonstrating its certain robustness to occlusions.

4.6 Robustness analysis against 2D input noise

Mainstream 2D detectors, such as Stacked Hourglass and HRNet, still output 17 keypoints when encountering occlusions, but the detection results of the occluded joints become inaccurate. To simulate the 2D keypoint detection errors caused by occlusions, we add zero-mean Gaussian noise with varying variances to the outputs of mainstream 2D detectors. We compare our model with state-of-the-art baselines, including MotionAGFormer and TCPFormer, and investigate the contribution of each proposed module to the noise robustness. All experiments uniformly use the 2D pose outputs obtained by the Stacked Hourglass detector on the Human3.6 M dataset, and exactly the same noise samples are used for all comparison models at each noise level to ensure a fair and consistent evaluation.

The experimental results are shown in Table 9. SBAHGNet achieves the best performance across all noise levels in the comparative experiments, and its advantage becomes more pronounced under high-noise conditions. Furthermore, the ablation experiments, presented in Table 10, reveal that the model with only the MSHFE module exhibits even better noise robustness than the full model, while the model with only the multi-scale skeletal bias module performs worse than the baseline. By analyzing the underlying mechanisms, we find that the low-pass sub-path of the MSHFE module effectively suppresses random Gaussian noise by performing local weighted averaging on the features aggregated by the GCN. In contrast, the multi-scale skeletal bias module relies heavily on the accuracy of input features, and its performance gain is significantly compromised under noise interference.

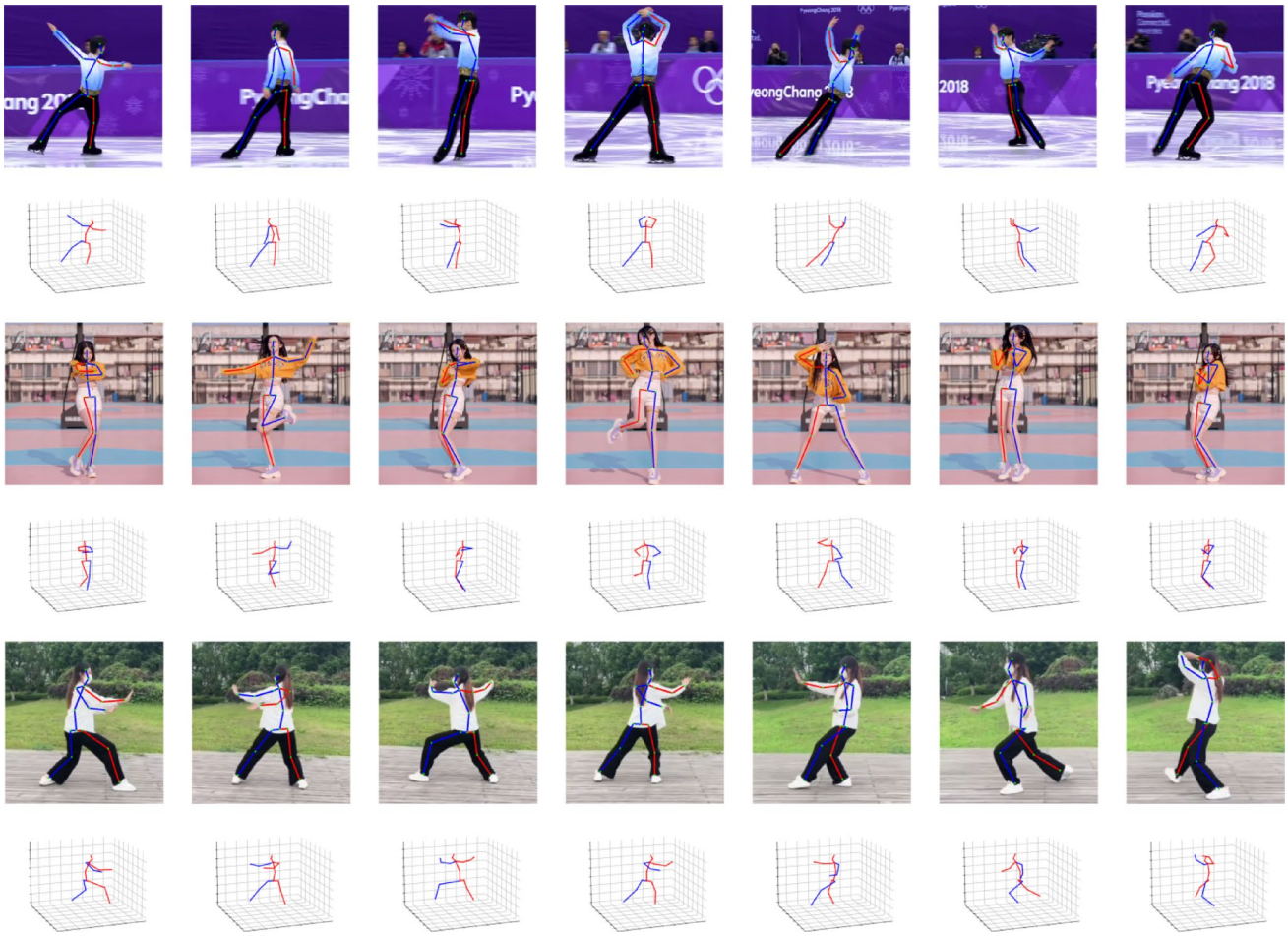


Fig. 7 Qualitative results of our method on in-the-wild videos

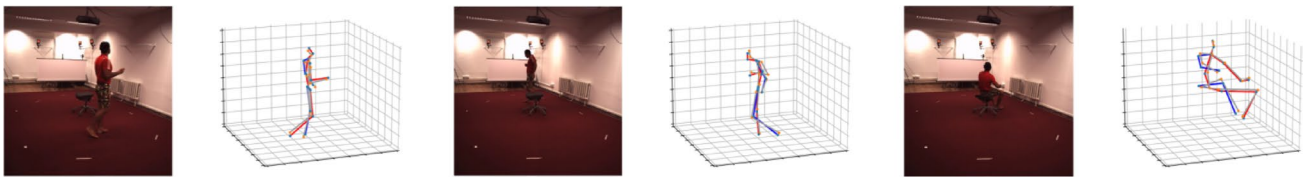


Fig. 8 Qualitative prediction results of SBAHGNet under occluded scenarios on the Human3.6 M dataset. The gray skeleton represents the ground-truth 3D pose, with the ground-truth keypoints shown in orange. Blue indicates the torso and left side, while red indicates the right side of the predicted pose

Table 9 MPJPE under zero-mean Gaussian noise with varying variance

MPJPE	σ^2 =4	σ^2 =25	σ^2 =64	σ^2 =100	σ^2 =225
MotionAG-Former-L [14]	*39.3*	*45.2*	*58.4*	*69.0*	*94.5*
TCPFormer [59]	39.5	55.2	75.8	88.6	117.5
SBAHGNet	**39.1**	**44.4**	**52.1**	**59.1**	**80.7**

The best results are highlighted in bold, and the second-best results are underlined

Table 10 Ablation study results of MPJPE under zero-mean Gaussian noise with varying variance

MPJPE	σ^2 =4	σ^2 =25	σ^2 =64	σ^2 =100	σ^2 =225
Baseline	40.8	59.8	85.7	101.7	136.0
MSHFE only	*39.2*	**42.6**	**48.1**	**52.7**	**66.3**
Multi-Scale Skeleton Bias only	40.5	63.5	89.7	103.9	132.1
SBAHGNet	**39.1**	*44.4*	*52.1*	*59.1*	*80.7*

The best results are highlighted in bold, and the second-best results are underlined

4.7 Complexity and inference efficiency analysis

We compare the inference latency of SBAHGNet and TCPFormer under both CPU and GPU hardware environments. All experiments are conducted under identical software and hardware settings to ensure fairness and reproducibility of the comparison results:

Hardware environment: Intel Core i7-12700KF CPU (3.61 GHz, 12 cores, 20 threads), 64 GB DDR4 RAM, NVIDIA GeForce RTX 3080 Ti GPU (12 GB VRAM).

Software environment: Microsoft Windows 11 Enterprise operating system, PyTorch 2.2.1 deep learning framework, CUDA 12.1 parallel computing platform, cuDNN 8.8.1 acceleration library.

Evaluation protocol: All models take a single sequence input with batch size = 1 (sequence length: 243 frames). Before measuring inference latency, we run 50 warm-up iterations to eliminate startup overhead caused by CUDA initialization and GPU cache. Then we record the average time over 100 consecutive inferences.

The experimental results are shown in Table 11. It can be observed that SBAHGNet achieves faster inference speed than TCPFormer on both GPU and CPU, with the advantage being more pronounced on CPU: GPU inference latency is reduced by 5.9% (66.7 ms vs. 70.9 ms), while CPU inference latency is reduced by 25.0% (1031.0 ms vs. 1375.3 ms).

Table 11 Comparison of parameter count, computational complexity, inference latency, and MPJPE (Protocol #1/#2) on the Human3.6 M dataset

Method	Param (M)	MACs (G)	GPU ms ↓	CPU ms ↓	P1↓/P2↓
TCPFormer	35.1	109.2G	70.9	1375.3	37.9/31.7
SBAHGNet	18.3	88.9G	66.7	1031.0	37.2/31.6

5 Conclusion

In this work, we propose SBAHGNet, a dual-branch spatio-temporal parallel architecture designed for 3D human pose estimation and motion prediction. The method consists of two complementary branches. The GCN branch applies explicit high-frequency compensation through MSHFE to the aggregated features, mitigating detail attenuation in fast and fine-grained motions caused by traditional graph convolutions. This, in turn, enhances dynamic realism in complex actions. The attention branch incorporates a multi-scale skeletal bias with only 376 learnable parameters into spatial self-attention, injecting flexible and powerful anatomical priors into the attention weights, which strengthens long-range joint dependencies and whole-limb coordinated modeling. The two branches operate in parallel with orthogonal strengths, achieving precise capture of local motion details and reasonable constraints on global pose structure without relying on complex temporal modelers. SBAHGNet achieves performance that is comparable to or exceeds current state-of-the-art methods on the Human3.6 M and MPI-INF-3DHP datasets, while maintaining high parameter and computational efficiency.

SBAHGNet still has clear limitations and room for optimization. The current model architecture defaults to the domain-general 243-frame long-term temporal clip as the standard input. Although it can adapt to video streams of arbitrary length through frame repetition or sliding window, the high computational overhead caused by long-term spatio-temporal modeling limits its direct deployment on edge devices. Meanwhile, the model still has room for improvement in terms of temporal smoothness and adaptability to multi-view scenarios. In future work, we plan to effectively reduce the computational burden of the model while maintaining estimation accuracy by adapting to dynamic short-sequence inputs, designing a lightweight network architecture, and introducing knowledge distillation technology, so as to improve its edge deployment capability. In addition, we will also explore the deep integration of SBAHGNet with existing mature pose optimization algorithms and multi-view aggregation methods, to further improve the model performance and scenario adaptability .Specifically, in monocular scenarios, we will attempt to deeply integrate SBAHGNet with SmoothNet [63]: first, we will adopt the native two-stage plug-and-play scheme of SmoothNet, take the high-precision pose predictions output by SBAHGNet as the input, and improve the smoothness of the output poses through its temporal optimization module. At the same time, we also plan to explore an end-to-end joint training scheme, with a view to achieving collaborative optimization between single-frame estimation accuracy and temporal smoothness. In multi-view scenarios, the monocular prediction results of each view can be attempted to be aggregated through the convex optimization method in COMETH [64], with the expectation of obtaining more consistent cross-view human pose estimation and tracking results.

Author contributions Yu Wang contributed to methodology, conceptualization, software, and writing-original draft; Jiaqiu Ai contributed to conceptualization, funding acquisition, resources, supervision, and writing-review and editing; Xinyu Sun helped with data curation, visualization, investigation, and writing-original draft; Yong Zhang contributed to resources and supervision; Jinyang Huang contributed to software and investigation; All authors have reviewed and approved thefinal manuscript.

Funding This work was supported by the National Natural Science Foundation of China under Grant 62471171, 62071164, by the Fundamental Research Funds for the Central Universities of China under Grant PA2025IISL0111.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Guo, Y., Gao, T., Dong, A., Jiang, X., Zhu, Z., Wang, F.: A survey of the state of the art in monocular 3D human pose estimation: methods, benchmarks, and challenges. Sensors. **25**(8), 2409 (2025). https://doi.org/10.3390/s25082409
2. Ben Gamra, M., Akhloufi, M.A.: A review of deep learning techniques for 2D and 3D human pose estimation. Image Vis. Comput. **114**, 104282 (2021). https://doi.org/10.1016/j.imavis.2021.104282
3. Peng, K., Yin, C., Zheng, J., Liu, R., Schneider, D., Zhang, J., Yang, K., Sarfraz, M.S., Stiefelhagen, R., Roitberg, A.: Navigatin-. **171**, 112239 (2026). https://doi.org/10.1016/j.patcog.2025.112239
4. Yang, A., Lu, W., Naeem, W., Chen, L., Fei, M.: A sequence models-based real-time multi-person action recognition method with monocular vision. J. Ambient Intell. Humaniz. Comput. (2021). https://doi.org/10.1007/s12652-021-03399-z
5. Cheng, Y., Yi, P., Liu, R., Dong, J., Zhou, D., Zhang, Q.: Hu-man-robot interaction method combining human pose estimati-on and motion intention recognition. In: 2021 IEEE 24th Inter-national Conference on Computer Supported Cooperative Wo-rk in Design (CSCWD), pp. 958–963 (2021). https://doi.org/10.1109/CSCWD49262.2021.9437772
6. Lin, H.-Y., Chen, T.-W.: Augmented reality with human body interaction based on monocular 3D pose estimation. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) Advanced Concepts for Intelligent Vision Systems. ACIVS 2010. Lecture Notes in Computer Science, vol. 6474, pp. 321–331. Springer, Berlin (2010)
7. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., Theobalt, C.: VNect: R-eal-time 3D human pose estimation with a single RGB came-ra. ACM Trans. Graph. **36**(4), 1–14 (2017). https://doi.org/10.1145/3072959.3073596
8. Sun, Z., Liang, Y., Ma, Z., Zhang, T., Bao, L., Li, G., He, S.: RePOSE: 3D human pose estimation via spatio-temporal depth relational consistency. In: Computer Vision—ECCV 2024. Lecture Notes in Computer Science, pp. 309–325. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72655-2_18
9. Liu, Y., Zhang, Z.: STGFormer: spatio-temporal GraphFormer for 3D human pose estimation in video. Pattern Recogn. **171**, 112239 (2026). https://doi.org/10.1016/j.patcog.2025.112239
10. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4929–4938 (2019). https://doi.org/10.1109/CVPR.2019.00794
11. Zhao, Q., Zheng, C., Liu, M., Chen, C.: A single 2D pose with c-ontext is worth hundreds for 3D human pose estimation. In: Advances in Neural Information Processing Systems 36NeurIPS, pp. 27394–27413 (2023) (2023)
12. Lee, S., Hwang, Y., Lee, J.T.: Learning 2D human poses for better 3D lifting via multi-model 3D-guidance. In: Proceedings of the Asian Conference on Computer Vision (ACCV) 2024. Lecture Notes in Computer Science, vol. 15472, pp. 185–202. Springer, Cham (2025). https://doi.org/10.1007/978-981-96-0885-0_11
13. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8818–8829 (2023)
14. Mehraban, S., Adeli, V., Taati, B.: Motionagformer: enhancing 3d human pose estimation with a transformer-gcnformer net-work. Mehraban, Adeli, S., Taati, V.: B.: Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network
15. Zhao, L., Peng, X., Tian, Y., Kapadia, M.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and p-attern recognition, pp. 3425–3435 (2019)
16. Zhang, J., Tu, Z., Yang, J., Yang, J., Chen, Y., Yuan, J.: MixS-TE: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13232–13242 (2022)
17. Chen, Z., Dai, J., Bai, J., Pan, J.: (or the published author list as in the journal): DGFormer: dynamic graph transformer for 3D human pose estimation. Pattern Recognit (2024). https://doi.org/10.1016/j.patcog.2024.110446
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (Poster) (2017). https://arxiv.org/abs/1609.02907
19. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 3844–3852 (2016)
20. Wu, F., de Souza, A.H. Jr., Zhang, T., Fifty, C., Yu, T., Weinberger, K.Q.: Simplifying graph convolutional networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, pp. 6861–6871 (2019). https://proceedings.mlr.press/v97/wu19e.html
21. Li, M., Chen, S., Zhang, Z., Xie, L., Tian, Q., Zhang, Y.: Skeleton-parted graph scattering networks for 3D human motion p-rediction. In: Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, Lecture Notes in Computer Science, pp. 18–36. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20068-7_2
22. Dong, Y., Ding, K., Jalaian, B., Ji, S., Li, J.: AdaGNN: graph neural networks with adaptive frequency response filter. In: Proceedings of the 30th ACM International Conference on I-nformation and Knowledge Management (CIKM), pp. 392–401 (2021). https://doi.org/10.1145/3459637.3482226
23. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding 208–209, 103219 (2021). https://doi.org/10.1016/j.cviu.2021.103219
24. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-tempor-al attention network for skeleton-based action-gesture recogniti-on. In: Computer Vision—ACCV 2020 (Asian Conference on Computer Vision), Lecture Notes in Computer Science, vol. 12626, pp. 38–53. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69541-5_3
25. Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W., Miao, Q.: Transformer for skeleton-based action recognition: a review of recent advances. Neurocomputing. **537**, 164–186 (2023). https://doi.org/10.1016/j.neucom.2023.03.001
26. Gao, Z., Wang, P., Lv, P., Jiang, X., Liu, Q., Wang, P., Xu, M., Li, W.: Focal and global spatial-temporal transformer for skeleton-based action recognition (FG-STFormer). In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 382–398 (2022). https://doi.org/10.1007/978-3-031-26316-3_10

27. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11656–11665 (2021)
28. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: exploring frequency domain for efficient and robust 3d hu-man pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8877–8886 (2023)
29. Jiang, Y., Sun, Z., Yu, S., Wang, S., Song, Y.: A graph skeleton transformer network for action recognition. Symmetry. **14**(8), 1547 (2022). https://doi.org/10.3390/sym14081547
30. Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W., Miao, Q.: Transformer for skeleton-based action recognition: a review of recent advances. Neurocomputing. **537**, 164–186 (2023). https://doi.org/10.1016/j.neucom.2023.03.001
31. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3D human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 68–84 (2018)
32. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., Theobalt, C.: VNect: real-time 3D human pose estimation with a single RGB came-ra. ACM Trans. Graph. **36**(4) (2017). https://doi.org/10.1145/3072959.3073596
33. You, Y., Liu, H., Wang, T., Li, W., Ding, R., Li, X.: Co-evolution of pose and mesh for 3D human body estimation from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14963–14973 (2023). https://doi.org/10.1109/ICCV51070.2023.01374
34. Zheng, K., et al.: 3D Human pose estimation via non-causal retentive networks. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024). https://doi.org/10.1007/978-3-031-73414-4_7
35. Hsu, C.-H., Jang, J.-S.R.: Enhancing 3D human pose estimation with bone length adjustment. In: Proceedings of the Asian Conference on Computer Vision (ACCV), Lecture Notes in Computer Science (LNCS), Springer (2024). https://doi.org/10.1007/978-981-96-0885-0_14
36. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 7444–7452 (2018). https://doi.org/10.1609/aaai.v32i1.12328
37. Azizi, N., Possegger, H., Rodolà, E., Bischof, H.: 3D Human pose estimation using möbius graph convolutional networks. In: Computer Vision—ECCV 2022, Proceedings, Part I, Lecture Notes in Computer Science, pp. 160–178. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19769-7_10
38. Zhang, Z.: Group Graph convolutional networks for 3D human pose estimation. In: Proceedings of the 33rd British Ma-chine Vision Conference (BMVC 2022). BMVA Press (2022)
39. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13156 (2022)
40. Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4790–4799 (2023)
41. Liu, J., Rojas, J., Li, Y., Liang, Z., Guan, Y., Xi, N., Zhu, H.: GAST-Net: graph attention spatio-temporal convolutional networks for 3D human pose estimation in video. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 3374–3380 (2021). https://doi.org/10.1109/ICRA48506.2021.9561605
42. Wang, T., Zhang, X.: Simplified-attention enhanced graph convolutional network for 3D human pose estimation (SaEGC-Net). Neurocomputing. **501**, 231–243 (2022). https://doi.org/10.1016/j.neucom.2022.06.033
43. Lin, H., Chiu, Y.-W., Wu, P.-Y.: AMPose: alternately mixed global-local attention model for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1–5 (2023)
44. Zhai, K., Nie, Q., Ouyang, B., Li, X., Yang, S.: HopFIR: Hop-wise GraphFormer with intragroup joint refinement for 3D human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14985–14995 (2023). https://doi.org/10.1109/ICCV51070.2023.01376
45. Ai, J., Mao, Y., Luo, Q., Jia, L., Xing, M.: SAR target classification using the multi-kernel-size feature fusion based convolutional neural network. IEEE Trans. Geosci. Remote Sens. **60**(5), 214313 (2022)
46. Ai, J., Fan, G., Mao, Y., Jin, J., Xing, M., Yan, H.: An improved SRGAN based ambiguity suppression algorithm for SAR ship target contrast enhancement. IEEE Geosci. Remote Sens. Lett. **19**(4), 4018705 (2022)
47. Huang, J., Feng, Y., Cui, F.-Q., Zhang, X., Liu, Z., Liu, X., Liu, J., Zhang, F., Li, M.: Identifying who you are no matter what you write through abstracting handwriting style. IEEE Trans. Depend. Secur. Comput. 1–15 (2026). https://doi.org/10.1109/TDSC.2026.3668275
48. Xue, W., Ai, J., Zhu, Y., Sun, X., Zhang, Y., Gao, G.: LMC-Net: light-weight modality compensation network for salient ship detection under missing modality conditions. IEEE Trans. Aerosp. Electron. Syst., pp. 1–14 (2026)
49. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2013)
50. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV), pp. 506–516 IEEE (2017)
51. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: PoseFormerV2: exploring frequency domain for efficient and robust 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8877–8886. IEEE (2023)
52. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: MotionBERT: a unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15085–15099. IEEE (2023)
53. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv:1711.05101 (2017)
54. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, pp. 483–499. Springer (2016)
55. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision, pp. 461–478. Springer (2022)
56. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: a unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15085–15099 (2023)
57. Chen, H., He, J.Y., Xiang, W., et al.: Hdformer: high-order directed transformer for 3d human pose Estimation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, pp. 581–589 (2023)

58. Qian, X., Tang, Y., Zhang, N., Han, M., Xiao, J., Huang, M., Lin, R.: Hstformer: hierarchical spatial-temporal transformers for 3d human pose estimation. (2023). https://doi.org/10.48550/arXiv.2301.07322
59. Liu, J., Liu, M., Liu, H., Li, W.: TCPFormer: learning temporal correlation with implicit pose proxy for 3D human pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (vol. 39) (2025). https://doi.org/10.1609/aaai.v39i5.32583
60. Foo, L.G., Li, T., Rahmani, H., Ke, Q., Liu, J.: Unified pose sequence modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 13019–13030 (2023). https://doi.org/10.1109/-CVPR52729.2023.01251
61. Farhadi, A., Redmon, J.: Yolov3: an incremental improvement. In: Computer Vision and Pattern Recognition, vol. 1804, pp. 1–6. Springer, Berlin/Heidelberg, Germany (2018)
62. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5693–5703 (2019)
63. Zeng, A., et al.: SmoothNet: A plug-and-play network for refining human poses in videos. (2021). https://doi.org/10.48550/arXiv.2112.13715 :2112.13715
64. Martini, E., et al.: COMETH: convex optimization for multiview estimation and tracking of humans. Expert Syst Appl, 210, 131728 (2026). https://doi.org/10.1016/j.eswa.2026.131728