

HSTGFormer: Hyper Spatial-Temporal Graph Transformer for 3D Human Pose Estimation

Ruochen Li
 ruochen.li@durham.ac.uk
 Shuang Chen
 shuang.chen@durham.ac.uk
 Wenke E
 wenke.e@durham.ac.uk
 Farshad Arvin
 farshad.arvin@durham.ac.uk
 Amir Atapour-Abarghouei
 amir.atapour-abarghouei@durham.ac.uk

Department of Computer Science
 Durham University
 Durham, UK

Abstract

Transformer-based methods have achieved strong performance in monocular 3D human pose estimation, but most existing approaches organise spatial and temporal reasoning as separate stages, which may weaken unified spatial-temporal interdependencies inherent in human motion and compress frame-level structural information before temporal modelling. In this paper, we propose HSTGFormer, a graph-enhanced Transformer framework that reformulates spatial-temporal reasoning as localised coupled graph aggregation over joint-time nodes. Specifically, HSTGFormer introduces a Hyper Spatial-Temporal Graph (HSTG), which decomposes global spatial-temporal reasoning into local spatial-temporal receptive fields around individual joint-time nodes by extending per-frame skeleton graphs into temporal neighbourhoods, thereby enabling structure-aware coupled reasoning while preserving local structural motion information. It further incorporates an Adaptive Dual-Scale Temporal Graph (ADSTG) to capture joint-specific temporal dependencies over complementary short- and long-range windows. A lightweight node-wise fusion module further adaptively integrates the two graph representations for each joint-time node. Experiments on Human3.6M and MPI-INF-3DHP show that HSTGFormer achieves strong accuracy with high computational efficiency.

1 Introduction

3D human pose estimation (HPE) aims to recover accurate human joint coordinates from input monocular 2D images, which is a crucial task in computer vision with broad research significance and related applications such as motion prediction [0, 65], human-computer interaction [0, 27, 28, 29], sports analysis [8, 60], and virtual reality [6]. Unlike single-frame pose estimation [08, 67, 89], video-based methods exploit both the anatomical structure of the human body and the temporal dynamics of motion. Skeleton provides strong kinematic

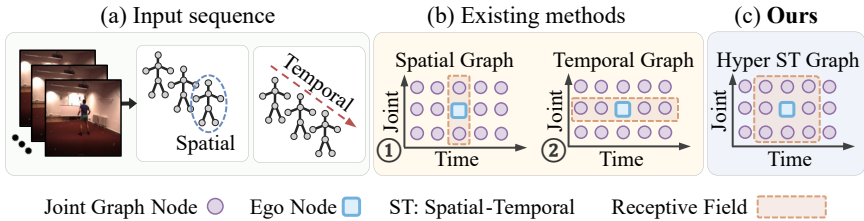


Figure 1: Illustration of the proposed *Hyper Spatial-Temporal Graph* (HSTG). (a) 2D joints are represented as graph nodes over time. (b) Existing methods decouple spatial and temporal reasoning through intra-frame skeletal aggregation followed by same-joint temporal modelling. (c) HSTG defines a localised coupled spatial-temporal receptive field around each ego node, efficiently aggregating from anatomically related joints within nearby frames.

constraints among connected joints, while consecutive frames contain motion cues that help resolve depth ambiguity, occlusion, and noisy detections. Therefore, effective 3D HPE requires comprehensive modelling of the spatial-temporal correlations of human motions.

Early methods for 3D human pose estimation typically operated on individual frames and adopted CNN-based architectures to either directly regress 3D joint coordinates from images [17] or to lift 2D keypoint detections to 3D poses using fully connected networks [22]. These methods generally lack effective and explicit spatial-temporal dependency modelling, relying primarily on per-frame predictions with limited structural priors. Inspired by the representation learning capability of Transformers [9, 33] in natural language processing, many video-based 3D HPE methods have adopted attention mechanisms to model spatial-temporal dependencies [20, 23, 26, 32, 38, 40, 42]. To model spatial-temporal correlations, these methods employ spatial attention to capture interactions among joints within each frame, while using temporal attention to model motion dependencies across consecutive frames. Despite their strong performance, Transformer-based methods often lack explicit skeletal priors and require stacked attention layers to capture complex spatial-temporal correlations, resulting in high computational cost for long video sequences. To address these limitations, recent works [13, 23, 26] incorporate graphs (Fig. 1(a)) into Transformer-based frameworks to introduce explicit skeletal priors and improve spatial-temporal representation learning. By modelling human joints as graph nodes, graph-based methods provide stronger structural inductive bias and enable efficient representation learning.

Despite recent progress, existing methods still face several limitations in spatial-temporal modelling. First, most existing methods organise spatial and temporal reasoning as two separate propagation stages, where spatial features are first extracted within each frame and subsequently passed to temporal modules for motion modelling (Fig. 1(b)). This separate process compresses frame-level structural information before temporal reasoning, limiting the ability to capture continuous spatial-temporal interdependencies [12, 13, 16, 21, 36]. For example, during walking, the knee and foot not only exhibit strong intra-frame anatomical coupling, but also maintain correlated local motion patterns across adjacent timesteps. However, sequential spatial-then-temporal propagation may weaken such localised spatial-temporal dependencies before coupled reasoning is established. Second, temporal modelling is often conducted over either long-range sequences or predefined receptive fields. Although recent works explore multi-scale temporal modelling [20], they still rely on fixed temporal scales, which may be insufficient for modelling joint dynamics across different motion patterns (e.g.,

fast-moving extremities vs. relatively stable torso joints).

In this paper, we propose **HSTGFormer**, a graph-based Transformer framework for efficient 3D human pose estimation. It introduces a graph-enhanced spatial-temporal reasoning paradigm based on two complementary graph structures: **Hyper Spatial-Temporal Graph** (HSTG) and **Adaptive Dual-Scale Temporal Graph** (ADSTG). As shown in Fig. 1(c), HSTG moves beyond decoupled per-frame spatial and per-joint temporal receptive fields by extending skeleton graphs into localised temporal neighbourhoods, yielding a local coupled spatial-temporal receptive field for each joint-time node. This formulation preserves localized spatial-temporal interdependencies within each joint-time neighbourhood, alleviating early information compression [21] of frame-level skeletal features before temporal modelling. Through efficient graph-based aggregation, HSTG preserves anatomical priors and captures local spatial-temporal context without constructing a dense spatial-temporal graph. For temporal modelling, ADSTG constructs dynamic graphs over short- and long-range temporal windows. This allows each joint to adaptively aggregate motion dependencies from different temporal contexts, enabling flexible temporal reasoning for diverse joint motion patterns. Finally, a context-aware fusion mechanism assigns node-wise weights to HSTG and ADSTG, allowing each node to integrate complementary spatial-temporal contexts.

Experiments on Human3.6M [9] and MPI-INF-3DHP [24] show that HSTGFormer achieves strong accuracy with high computational efficiency. Our contributions are:

- We propose **HSTGFormer**, a graph-based Transformer framework for efficient 3D human pose estimation, which reformulates spatial-temporal modelling from a graph perspective to better capture motion continuity and structural dependencies.
- We introduce a *Hyper Spatial-Temporal Graph* (HSTG) structure (Sec. 3.2) that extends per-frame skeleton graphs to localised temporal neighbourhoods, enabling structure-aware modelling of local motion continuity while preserving human skeletal priors.
- We design an *Adaptive Dual-Scale Temporal Graph* (ADSTG) module (Sec. 3.3) to construct and fuse dynamic temporal graphs over complementary temporal ranges, allowing joint-specific and adaptive temporal dependency modelling.

2 Related Work

2.1 Transformer for 3D Human Pose Estimation

With the success of Transformers [53] in natural language processing [9] and computer vision [9, 14], recent works have increasingly adopted Transformer-based architectures for 3D human pose estimation. PoseFormer [40] introduces one of the first pure Transformer frameworks for 3D pose lifting, where spatial and temporal Transformers are stacked to jointly model spatial relationships within each frame and long-range dependencies across frames. MixSTE [58] proposes a mixed spatial-temporal encoder with an alternating seq2seq design that more effectively captures temporal dynamics of different body joints over long sequences. MotionBERT [42] employs large-scale pre-training to learn general motion representations from massive pose data and adapts them to downstream 3D pose estimation tasks. STCFormer [52] introduces a two-pathway design for spatial-temporal modelling, enabling more expressive modeling of complex motion patterns in monocular videos. TCPFormer [20] further refines temporal context aggregation by tailoring Transformer blocks to better capture

multi-scale temporal cues and mitigate motion ambiguity, leading to consistent performance gains on standard 3D pose benchmarks. However, most Transformer-based methods do not explicitly exploit human skeletal priors, often requiring deeper architectures to model complex spatial-temporal dependencies, which increases computational cost.

2.2 Graph-transformer for 3D Human Pose Estimation

Due to the strong capability for modelling topological structures, graph-based methods have been widely adopted for spatial-temporal modelling in 3D human pose estimation. PoseG-TAC [14] combines graph atrous convolution for multi-scale neighbour aggregation with graph Transformer layers to capture long-range spatial-temporal dependencies. MotionAGFormer [23] integrates attention and GCN [10, 11] branches to capture global joint relationships and local skeletal dependencies. DiffPose [5] further introduces a diffusion-based framework with GCN to better encode spatial dependencies on the human skeleton and generate more reliable multi-hypothesis 3D poses under occlusion. GLA-GCN [67] explores adaptive graph convolution to capture global pose representations. By adopting a strided temporal design, it reduces the effective temporal modelling range and achieves competitive performance compared with Transformer-based methods, while requiring lower memory consumption. KTPFormer [26] introduces kinematic and trajectory priors into spatial and temporal attention to better encode anatomical and motion constraints. However, most existing graph-transformer methods still perform spatial modelling within individual frames, limiting their ability to explicitly capture continuous spatial-temporal interdependencies during human motion.

3 Method

3.1 Problem Formulation and Overview

Given an input 2D pose sequence $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T \in \mathbb{R}^{T \times J \times C_{in}}$, where T is the number of frames and J is the number of body joints, each joint is represented by its 2D coordinates and confidence score. The objective is to recover the corresponding 3D pose sequence $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^T \in \mathbb{R}^{T \times J \times C_{out}}$, where C_{in} and C_{out} denote the input and output joint dimensions, respectively. Following recent Transformer-based 3D pose estimation methods [20, 23], we first map the input sequence into a high-dimensional latent space through a linear embedding layer, producing $\mathbf{X}_h \in \mathbb{R}^{T \times J \times D}$, where D is the hidden feature dimension. To retain joint identity information, a learnable joint-level positional encoding $\mathbf{P}_{pos}$ is added to the embedded features. The resulting representation is then processed by a spatial-temporal Transformer encoder $\mathbf{X}_{st} = \Phi(\mathbf{X}_h + \mathbf{P}_{pos})$, where $\Phi(\cdot)$ denotes the backbone encoder used to capture global spatial-temporal dependencies before graph-based reasoning.

As illustrated in Fig. 2, building on the embedded pose representation $\mathbf{X}_h$, HSTGFormer performs graph-enhanced spatial-temporal reasoning through two complementary branches. The first branch, *Hyper Spatial-Temporal Graph* (HSTG), extends the per-frame skeleton graph to localised temporal neighbourhoods, forming a coupled spatial-temporal receptive field for each joint-time node. Through mask-constrained and factorised graph attention, HSTG captures local structural motion context while preserving anatomical connectivity priors. The second branch, *Adaptive Dual-Scale Temporal Graph* (ADSTG), constructs dynamic graphs over short- and long-range temporal windows, allowing each joint to adaptively aggregate motion cues from complementary temporal ranges. A context-aware node-wise

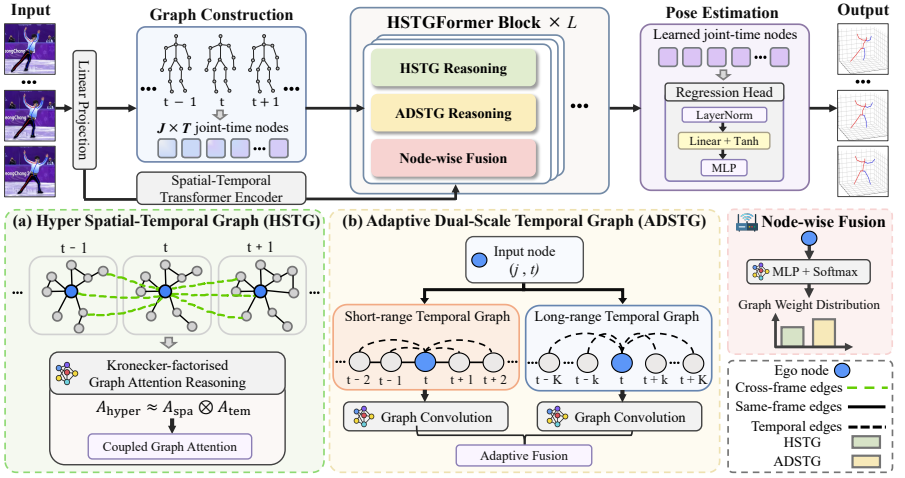


Figure 2: Framework overview. HSTGFormer consists of two graph-based modules: (a) the Hyper Spatial-Temporal Graph (HSTG), which constructs localised coupled spatial-temporal receptive fields across nearby frames, and (b) the Adaptive Dual-Scale Temporal Graph (ADSTG), which models adaptive short- and long-range temporal dependencies. A node-wise fusion module then adaptively integrates graph contexts for final 3D pose estimation.

fusion module then predicts adaptive weights for the two branches, enabling each joint-time node to integrate local spatial-temporal context and temporal motion dynamics. The fused graph representation is finally combined with the backbone features and fed into the regression head to estimate the 3D pose sequence.

3.2 Hyper Spatial-Temporal Graph (HSTG) Reasoning

Human motion is governed by anatomical constraints and local temporal continuity, but conventional spatial-then-temporal methods model them separately: first within each frame, then along each joint trajectory. Such designs may overlook localised spatial-temporal interdependencies, where neighbouring joints exhibit correlated motion over adjacent timesteps. To address this, we introduce the *Hyper Spatial-Temporal Graph* (HSTG), which extends the skeleton graph from individual frames to local temporal neighbourhoods and performs structure-aware graph reasoning over joint-time nodes.

Graph Construction. Given the embedded pose representation $\mathbf{X}_h \in \mathbb{R}^{T \times J \times D}$, we construct a Hyper Spatial-Temporal Graph $\mathcal{G}_{\text{hyper}} = (\mathcal{V}, \mathcal{E}_{\text{hyper}})$, where each node $v_{t,j} \in \mathcal{V}$ represents joint j at timestep t , and $|\mathcal{V}| = TJ$. From the perspective of each ego node $v_{t,j}$, HSTG defines a local spatial-temporal neighbourhood by jointly considering anatomical connectivity and temporal proximity. Specifically, let $\mathbf{A}_{\text{spsa}} \in \{0, 1\}^{J \times J}$ denote the skeleton adjacency matrix with self-loops, and let $\mathbf{A}_{\text{tem}} \in \{0, 1\}^{T \times T}$ denote a temporal band adjacency:

$$[\mathbf{A}_{\text{tem}}]_{t,t'} = \begin{cases} 1, & |t - t'| \leq w, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where w is the temporal window radius. The resulting spatial-temporal support can be represented in a Kronecker-factorised form:

$$\mathbf{A}_{\text{hyper}} \approx \mathbf{A}_{\text{tem}} \otimes \mathbf{A}_{\text{spa}}, \quad (2)$$

where $\otimes$ denotes the Kronecker product. Here, $\mathbf{A}_{\text{hyper}}$ denotes the induced local spatial-temporal support rather than an explicitly materialised dense adjacency matrix. This factorised formulation allows HSTG to perform graph attention through spatial and temporal graph operations while preserving a localised coupled spatial-temporal receptive field.

Factorised Graph Attention. The support $\mathbf{A}_{\text{hyper}}$ specifies the valid local spatial-temporal neighbourhood of each joint-time node. However, explicitly performing attention over $\mathbf{A}_{\text{hyper}}$ would require constructing a $(TJ) \times (TJ)$ attention map, which is computationally expensive for long pose sequences. Instead, we exploit the Kronecker structure of $\mathbf{A}_{\text{hyper}}$ and implement HSTG reasoning through two factorised masked graph attention [64] operations: skeleton-constrained spatial attention followed by local temporal attention.

Given a set of node features $\mathbf{F} \in \mathbb{R}^{N \times D}$ and a binary adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$, we define a generic masked graph attention operator [63, 64] as

$$\begin{aligned} \text{GraphAttn}(\mathbf{F}, \mathbf{A})_i &= \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{F}_j \mathbf{W}_v, \\ \alpha_{ij} &= \frac{\exp((\mathbf{F}_i \mathbf{W}_q)(\mathbf{F}_j \mathbf{W}_k)^\top / \sqrt{D})}{\sum_{k \in \mathcal{N}(i)} \exp((\mathbf{F}_i \mathbf{W}_q)(\mathbf{F}_k \mathbf{W}_k)^\top / \sqrt{D})}, \end{aligned} \quad (3)$$

where $\mathcal{N}(i) = \{j \mid \mathbf{A}_{ij} = 1\}$ denotes the masked neighbourhood of node i , and $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable projection matrices.

Using this operator, we first normalise [11] the embedded feature as $\tilde{\mathbf{X}}_h = \text{LN}(\mathbf{X}_h)$. For each timestep t , we apply skeleton-constrained graph attention over the joint dimension:

$$\mathbf{Z}_t = \text{GraphAttn}_{\text{spa}}(\tilde{\mathbf{X}}_h^t, \mathbf{A}_{\text{spa}}), \quad (4)$$

where $\tilde{\mathbf{X}}_h^t \in \mathbb{R}^{J \times D}$ denotes the joint features at timestep t . Then, for each joint j , temporal graph attention is applied over its local temporal neighbourhood:

$$\mathbf{H}^j = \text{GraphAttn}_{\text{tem}}(\mathbf{Z}^j, \mathbf{A}_{\text{tem}}), \quad (5)$$

where $\mathbf{Z}^j \in \mathbb{R}^{T \times D}$ denotes the temporal sequence of the spatially enriched features for joint j . Finally, we apply residual refinement to obtain the HSTG output:

$$\mathbf{X}_{\text{hyper}} = \mathbf{X}_h + \sigma(\text{LN}(\mathbf{H} + \tilde{\mathbf{X}}_h \mathbf{W}_u)), \quad (6)$$

where $\sigma(\cdot)$ denotes the ReLU activation.

HSTG defines local joint-time receptive fields instead of sequentially separating spatial and temporal modelling. Its factorised attention preserves the localised Cartesian support between skeleton and temporal neighbours, enabling efficient spatial-temporal graph reasoning without constructing a dense attention graph.

3.3 Adaptive Dual-Scale Temporal Graph (ADSTG) Reasoning

Although HSTG captures localised coupled spatial-temporal context, human motion also involves temporal dependencies at different ranges. For example, fast-moving joints require short-range cues for fine-grained motion changes, while more stable joints may benefit from longer-range context. In this paper, we propose an *Adaptive Dual-Scale Temporal Graph* (ADSTG), which constructs dynamic temporal graphs over complementary short- and long-range windows and fuses them in a joint-aware manner.

Dynamic Temporal Graph Construction. Given the embedded pose representation $\mathbf{X}_h \in \mathbb{R}^{T \times J \times D}$, ADSTG constructs dynamic temporal graphs along the temporal dimension. Unlike HSTG, which defines a localised coupled spatial-temporal receptive field, ADSTG focuses on selecting informative temporal neighbours for each joint-time node (t, j) from complementary temporal ranges. Specifically, we define two centred temporal windows:

$$\mathcal{N}_m(t) = \{t' \mid |t - t'| \leq r_m\}, \quad m \in \{\text{short}, \text{long}\}, \quad (7)$$

where r_m denotes the temporal radius of branch m . Within each window, ADSTG computes dot-product similarities between node (t, j) and its temporal candidates (t', j) , and retains the top- k_m most similar candidates to form a sparse dynamic temporal adjacency:

$$[\mathbf{A}_m]_{t,t'} = \begin{cases} 1, & t' \in \text{TopK}_{k_m}(\mathcal{N}_m(t)), \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where $\mathbf{A}_{\text{short}}$ and $\mathbf{A}_{\text{long}}$ denote the short- and long-range dynamic temporal adjacencies, respectively. This construction allows each joint-time node to adaptively select relevant temporal neighbours within local and extended temporal contexts.

Dual-Scale Temporal Propagation. Given a dynamic temporal adjacency $\mathbf{A}$, we define a normalised temporal graph GCN [14] as

$$\begin{aligned} \text{TempGCN}(\mathbf{X}_h, \mathbf{A}) &= \tilde{\mathbf{A}} \mathbf{X}_h \mathbf{W}_v, \\ \tilde{\mathbf{A}} &= \mathbf{D}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{D}^{-\frac{1}{2}}, \end{aligned} \quad (9)$$

where self-loops are added before symmetric normalisation, $\mathbf{D}$ is the degree matrix, and $\mathbf{W}_v$ is a learnable projection matrix. ADSTG applies the same temporal GCN to the short- and long-range dynamic temporal adjacencies:

$$\mathbf{Z}_{\text{short}} = \text{TempGCN}(\mathbf{X}_h, \mathbf{A}_{\text{short}}), \quad \mathbf{Z}_{\text{long}} = \text{TempGCN}(\mathbf{X}_h, \mathbf{A}_{\text{long}}). \quad (10)$$

This produces two complementary temporal representations, corresponding to local motion details and extended temporal context.

To combine the two temporal scales, we use a lightweight topology-aware gate based on the skeleton degree. Given $\mathbf{A}_{\text{spsa}}$, we compute the joint degree vector $\mathbf{d} \in \mathbb{R}^J$ and generate joint-wise scale weights $\mathbf{G}$. The output of ADSTG is obtained by weighted fusion of the short- and long-range temporal representations:

$$\mathbf{X}_{\text{ada}} = \sum_{m \in \{\text{short}, \text{long}\}} \mathbf{G}_m \odot \mathbf{Z}_m. \quad (11)$$

ADSTG complements HSTG by focusing on adaptive temporal dependency modelling. Instead of relying on a single fixed temporal neighbourhood, ADSTG constructs content-adaptive temporal graphs within short- and long-range windows, allowing each joint to select informative temporal neighbours according to its motion pattern. The joint-adaptive fusion further balances local and extended temporal contexts, providing flexible temporal reasoning for heterogeneous human motion dynamics.

3.4 Adaptive Node-wise Fusion

After obtaining the HSTG feature $\mathbf{X}_{\text{hyper}}$ and the ADSTG feature $\mathbf{X}_{\text{ada}}$, we use two lightweight branch-specific Multilayer Perceptrons (MLP) for feature enhancement:

$$\hat{\mathbf{X}}_{\text{hyper}} = \text{MLP}_{\text{hyper}}(\mathbf{X}_{\text{hyper}}), \quad \hat{\mathbf{X}}_{\text{ada}} = \text{MLP}_{\text{ada}}(\mathbf{X}_{\text{ada}}). \quad (12)$$

To adaptively combine the two graph representations, we introduce a context-aware fusion module that predicts node-wise weights for each joint-time node (t, j) . Here, $\mathbf{C}_{\text{spa}}$ is obtained by averaging $\mathbf{X}_h$ over joints to provide frame-level context, while $\mathbf{C}_{\text{tem}}$ is obtained by averaging over time to provide joint-level context. After broadcasting them to each joint-time node, we concatenate the token feature with $\mathbf{C}_{\text{spa}}$ and $\mathbf{C}_{\text{tem}}$, and feed the result into a lightweight MLP followed by softmax:

$$\omega_{\text{hyper}}, \omega_{\text{ada}} = \text{Softmax} \left(\text{MLP} \left[\mathbf{X}_h, \mathbf{C}_{\text{spa}}, \mathbf{C}_{\text{tem}} \right] \right). \quad (13)$$

The final graph representation is then computed as

$$\mathbf{X}_{\text{fuse}} = \omega_{\text{hyper}} \odot \hat{\mathbf{X}}_{\text{hyper}} + \omega_{\text{ada}} \odot \hat{\mathbf{X}}_{\text{ada}}. \quad (14)$$

This adaptive fusion allows each joint-time node to balance local coupled spatial-temporal reasoning from HSTG and adaptive temporal dependency modelling from ADSTG. The fused representation is then combined with the backbone Transformer encoder features [20, 23], enabling the model to retain global spatial-temporal context while incorporating graph-enhanced motion representations. After L stacked HSTGFormer blocks, an MLP-based regression head maps the learned joint-time features to the final 3D pose sequence.

3.5 Overall Learning Objectives

We optimise HSTGFormer end-to-end using a composite training objective following the commonly used protocol in recent 3D HPE methods [20]. The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{mpjpe}} + \lambda_s \mathcal{L}_{\text{nmjppe}} + \lambda_v \mathcal{L}_{\text{vel}} + \lambda_d \mathcal{L}_{\text{diff}} + \lambda_{lb} \mathcal{L}_{lb}, \quad (15)$$

where $\mathcal{L}_{\text{mpjpe}}$ and $\mathcal{L}_{\text{nmjppe}}$ measure the pose reconstruction error using MPJPE and normalised MPJPE, respectively. $\mathcal{L}_{\text{vel}}$ supervises motion consistency through velocity differences, while $\mathcal{L}_{\text{diff}}$ regularises consecutive-frame variations to encourage temporally smooth predictions [20]. We set $\lambda_s = 0.5$, $\lambda_v = 20$, $\lambda_d = 0.5$ in all experiments. Since our fusion module assigns node-wise weights to different graph pathways, we further apply a lightweight load-balancing regularisation to avoid one pathway dominating the fusion process:

$$\mathcal{L}_{lb} = \sum_{r=1}^K \left(\frac{1}{|\Omega|} \sum_{i \in \Omega} \omega_r^i - \frac{1}{K} \right)^2, \quad (16)$$

where K denotes the number of graph pathways, Ω is the set of all joint-time nodes in a mini-batch, and ω_r^i is the fusion weight assigned to pathway r for node i .

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate the proposed HSTGFormer on two standard benchmarks for monocular 3D human pose estimation: Human3.6M [9] and MPI-INF-3DHP [24]. These two datasets provide complementary evaluation settings, with Human3.6M mainly focusing on controlled indoor scenarios and MPI-INF-3DHP covering more diverse indoor and outdoor environments.

Human3.6M is a widely used 3D human pose estimation benchmark with 3.6M annotated frames from 11 subjects performing 15 actions in indoor settings. Following [15, 20, 23], we train on S1, S5, S6, S7, and S8, and evaluate on S9 and S11. We report Mean Per-Joint Position Error (MPJPE) and Procrustes-aligned MPJPE (P-MPJPE) as evaluation metrics.

MPI-INF-3DHP is a more challenging benchmark containing diverse indoor and outdoor motions with larger variations in viewpoints, poses, and environmental conditions, making it suitable for evaluating the generalisation ability. Following [15, 20, 23], we report MPJPE, Percentage of Correct Keypoints (PCK), and Area Under the Curve (AUC).

4.1.2 Implementation Details

Our framework is implemented in PyTorch and trained end-to-end using AdamW with a weight decay of 0.01. The feature dimension is set to 128.

Human3.6M: We train for 80 epochs with batch size 6 and sequence length 243. The initial learning rate is 5×10^{-4} with an exponential decay factor of 0.99. The short/long ADSTG windows are set to 27/81 frames. Following [20, 23], we evaluate with both ground-truth 2D keypoints and Stacked Hourglass detections [23].

MPI-INF-3DHP: We train for 90 epochs with batch size 6 and sequence length 81. The initial learning rate and decay follow Human3.6M. The short/long ADSTG windows are set to 9/27 frames. Following [20, 23], we use ground-truth 2D inputs for evaluation.

4.2 Quantitative Comparisons with State-of-the-Art Methods

In this section, we quantitatively compare the proposed HSTGFormer with recent state-of-the-art methods on Human3.6M and MPI-INF-3DHP. The comparisons include a broad range of Transformer-based and graph-enhanced pose estimation approaches. Notably, MotionAGFormer [23] and TCPFormer [20] use the same backbone encoder as our method, enabling a fair comparison of the proposed graph reasoning modules.

Indoor Monocular Human3.6M Dataset. As shown in Table 1, our method achieves highly competitive performance on Human3.6M, obtaining the best P-MPJPE of 31.5 mm and matching the best MPJPE of 37.9 mm among all compared methods. Compared with KTPFormer, our model reduces MPJPE from 40.1 mm to 37.9 mm and P-MPJPE from 31.9 mm to 31.5 mm, corresponding to relative improvements of 5.5% and 1.3%, respectively. Compared with MotionAGFormer-L, which shares the same backbone encoder, our method improves MPJPE from 38.4 mm to 37.9 mm and P-MPJPE from 32.5 mm to 31.5 mm. These results show that HSTGFormer improves pose estimation mainly through more effective graph-enhanced spatial-temporal reasoning, rather than simply increasing model capacity.

Efficiency Comparison. Beyond accuracy, our method achieves a favourable accuracy-efficiency trade-off. Compared with TCPFormer, which uses the same backbone encoder, our

Method	Venue	CE	T	Parameter	MACs	MACs/frames↓	MPJPE↓	P-MPJPE↓
MHFormer [13]	CVPR'22	✓	351	30.9M	7.1G	7096M	43.0	34.4
MixSTE [13]	CVPR'22	✗	243	33.6M	139.0G	572M	40.9	32.6
P-STMO [14]	ECCV'22	✓	243	6.2M	0.7G	740M	42.8	34.4
PoseFormerV2 [15]	CVPR'23	✓	243	14.3M	0.5G	528M	45.2	35.6
GLA-GCN [16]	ICCV'23	✓	243	1.3M	1.5G	1556M	44.4	34.8
MotionBERT [17]	ICCV'23	✗	243	42.3M	174.8G	719M	39.2	32.9
MotionAGFormer-L [18]	WACV'24	✗	243	19.0M	78.3G	322M	38.4	32.5
PoseRetNet [19]	ECCV'24	✗	243	25.2M	104.5G	430M	40.4	32.5
KTPFormer [20]	CVPR'24	✗	243	33.7M	69.5G	<u>286M</u>	40.1	31.9
H2OT+MotionAGFormer [21]	TPAMI'25	✗	243	11.7M	38.9G	-	<u>38.5</u>	-
TCPFormer [22]	AAAI'25	✗	243	35.1M	109.2G	449M	37.9	<u>31.7</u>
Ours	-	✗	243	14.2M	58.5G	240M	37.9	31.5

Table 1: Results on Human3.6M in millimeters under MPJPE and P-MPJPE. CE indicates centre-frame prediction. T is the number of input frames. MACs/frame represents the number of multiply-accumulate operations per output frame, where a lower value indicates higher computational efficiency. **Bold/underline** indicates the best/second-best result. Blue rows indicate methods sharing the same backbone encoder.

method obtains comparable MPJPE and better P-MPJPE with 59.5% fewer parameters and 46.5% lower MACs/frame. It also reduces parameters and MACs/frame by 25.3% and 25.5% compared with MotionAGFormer-L.

In-the-wild MPI-INF-3DHP Datasets. Table 2 reports the quantitative results on MPI-INF-3DHP, which contains more diverse scenes and larger motion variations than Human3.6M. Our method achieves the best overall performance. Compared with TCPFormer, which shares the same backbone encoder, our method improves AUC from 87.7 to 89.3 and reduces MPJPE from 15.0 mm to 14.0 mm (-6.7%). Compared with MotionAGFormer-L, our method improves AUC by 4.0 points and reduces MPJPE by 13.6%. These gains on the more challenging MPI-INF-3DHP benchmark indicate that the proposed HSTG and ADSTG modules improve robustness to diverse motion patterns and scene variations.

4.3 Ablation Studies and Analysis

4.3.1 Main Component Ablation Studies

Table 3 presents the main ablation study on Human3.6M. We evaluate the contribution of the two proposed graph reasoning modules, HSTG and ADSTG, as well as the adaptive node-wise fusion strategy. Variants (1)-(3) mainly examine the effectiveness of the graph reasoning design. Compared with the baseline without HSTG and ADSTG, adding either component brings clear improvements, reducing MPJPE from 41.1 mm to 39.2 mm and 38.8 mm, respectively. This shows that both local coupled spatial-temporal reasoning and adaptive temporal graph modelling contribute to better pose estimation. Variants (4)-(5) further analyse the fusion mechanism. Replacing node-wise adaptive fusion with fixed 0.5/0.5 weights degrades MPJPE from 37.9 mm to 38.3 mm, indicating that different joint-time nodes benefit from different graph contexts. Removing the load-balance loss also weakens the performance, suggesting that balanced utilization of the two branches helps stabilize adaptive fusion. Overall, the full HSTGFormer achieves the best MPJPE and P-MPJPE, showing that HSTG and ADSTG provide complementary spatial-temporal representations, while the proposed

Method	Venue	CE	T	PCK↑	AUC↑	MPJPE↓
MHFormer [15]	CVPR'22	✓	9	93.8	63.3	58.0
MixSTE [25]	CVPR'22	✗	27	94.4	66.5	54.9
P-STMO [26]	ECCV'22	✓	81	97.9	75.8	32.2
PoseFormerV2 [25]	CVPR'23	✓	81	97.9	78.8	27.8
GLA-GCN [25]	ICCV'23	✓	81	98.5	79.1	27.7
MotionBERT [25]	ICCV'23	✗	-	-	-	-
MotionAGFormer-L [25]	WACV'24	✗	81	98.2	85.3	16.2
PoseRetNet [25]	ECCV'24	✗	81	99.1	84.4	22.2
KTPFormer [25]	CVPR'24	✗	81	98.9	85.9	16.7
H2OT+MotionAGFormer [25]	TPAMI'25	✗	81	99.1	85.2	18.0
TCPFormer [25]	AAAI'25	✗	81	<u>99.0</u>	<u>87.7</u>	<u>15.0</u>
Ours	—	✗	81	99.1	89.3	14.0

Table 2: Results on MPI-INF-3DHP. CE indicates centre-frame prediction. T denotes the number of input frames. **Bold/underline** indicates the best/second-best result. Blue rows indicate methods sharing the same backbone encoder.

Variant	MPJPE↓	P-MPJPE↓
(1) Baseline (w/o HSTG & ADSTG)	41.1	36.0
(2) w/o HSTG	39.2	32.9
(3) w/o ADSTG	38.8	<u>32.5</u>
(4) w/o Node-wise Adaptive Fusion (fixed 0.5/0.5 weights)	<u>38.3</u>	32.6
(5) w/o load-balance loss ($\mathcal{L}_{lb}$)	38.8	33.0
HSTGFormer (Ours)	37.9	31.5

Table 3: Main component ablation on Human3.6M. We evaluate the contribution of the proposed HSTG, ADSTG, adaptive fusion strategy, and load-balance loss. Results are reported in MPJPE and P-MPJPE (mm). **Bold/underline** indicates the best/second-best result.

node-wise fusion further improves adaptive spatial-temporal reasoning.

4.3.2 Graph Design Analysis

Table 4 explores the design choices of HSTG and ADSTG by replacing them with alternative architectures. For HSTG, factorised MLPs lead to a clear performance drop, with our design reducing MPJPE by 4.5% and P-MPJPE by 4.2%, indicating that point-wise transformations are insufficient for modelling structured joint dependencies and coupled spatial-temporal interactions. Factorised GCNs perform better than MLPs but still underperform our design, suggesting that adaptive graph attention enables more flexible spatial-temporal aggregation. For ADSTG, replacing adaptive temporal graph modelling with fixed dual-scale temporal convolutions or fixed temporal attention also degrades performance. This shows that content-adaptive temporal neighbourhoods are more effective than predefined temporal kernels.

4.3.3 Per-action Quantitative Analysis

To better understand the action-wise behaviour of HSTGFormer, we analyse the per-action performance on Human3.6M in Table 5. Our method achieves the best results on several representative actions, including *Direction*, *Eating*, *Purchases*, *Sitting*, *Sitting Down*, *Walking*

Component	Variant	MPJPE↓	P-MPJPE↓
HSTG	(1) w/ Factorised MLP	39.7	33.0
	(2) w/ Factorised GCN	<u>38.1</u>	<u>32.1</u>
	Ours	37.9	31.5
ADSTG	(1) w/ Fixed Dual-Scale Temporal Convolution	39.1	32.7
	(2) w/ Fixed Temporal Attention	<u>38.7</u>	<u>32.6</u>
	Ours	37.9	31.5

Table 4: Ablation studies on HSTG and ADSTG design choices. Results are reported in MPJPE and P-MPJPE (mm). **Bold/underline** indicate the best/second-best results.

Method	Venue	T	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.
MHFormer [12]	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
MixSTE [13]	CVPR'22	243	32.0	34.2	31.7	33.7	34.4	39.2	32.0	31.8	42.9	46.9	35.5	32.0	34.4	23.6	25.2
P-STMO [14]	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
GLA-GCN [15]	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
MotionAGFormer-L [16]	WACV'24	243	31.0	32.6	<u>31.1</u>	28.0	34.0	38.8	31.5	30.1	41.4	<u>45.5</u>	34.9	30.8	31.3	22.8	23.2
KTPFormer [17]	CVPR'24	243	<u>30.1</u>	32.3	29.6	30.8	32.3	37.3	30.0	<u>30.2</u>	<u>41.0</u>	45.3	33.6	<u>29.9</u>	31.4	<u>21.5</u>	22.6
PoseRetNet [18]	ECCV'24	243	30.8	33.1	31.3	31.8	33.4	37.7	30.1	30.5	43.4	<u>45.5</u>	<u>34.3</u>	30.3	31.5	21.4	22.7
TCPFormer [19]	AAAI'25	243	<u>30.1</u>	31.6	31.4	27.3	33.5	<u>37.6</u>	29.4	29.6	41.1	45.9	34.4	29.6	<u>30.6</u>	21.7	<u>22.3</u>
Ours	-	243	29.9	<u>31.9</u>	29.6	<u>27.7</u>	<u>33.3</u>	<u>37.6</u>	<u>29.5</u>	29.3	39.6	45.3	34.8	30.0	30.2	22.3	22.1

Table 5: Results on Human3.6M under P-MPJPE for 15 actions. **Bold/underline** indicates the best/second-best result. Blue rows indicate methods sharing the same backbone encoder.

Dog, and *Walking Together*, while remaining competitive on *Discussion*, *Phone*, *Photo*, and *Pose*. The gains are particularly clear on complex or motion-intensive actions, suggesting that HSTGFormer effectively captures continuous spatial-temporal dependencies.

4.4 Qualitative Analysis

4.4.1 Pose Estimation Visualisation

Figures 3 and 4 present qualitative comparisons on both in-the-wild videos and the Human3.6M dataset. Compared with MotionAGFormer and TCPFormer, our method produces more structurally consistent and anatomically plausible 3D poses under challenging body articulations and self-occlusions. In particular, our predictions better preserve local joint geometry and long-range body coordination in difficult regions such as arms, shoulders, and torso bending poses. These qualitative results further demonstrate the effectiveness of the proposed HSTG and ADSTG modules for robust spatial-temporal pose reasoning.

4.4.2 Correlation Weights Analysis for HSTG

Figure 5 (a) visualizes the learned spatial-temporal correlation weights in HSTG. For each query joint-time node, we show its correlation responses to anatomically related joints within a local temporal neighbourhood. The highlighted regions indicate that the most relevant information does not always come from the same frame. Instead, the query node can attend to related joints across nearby timesteps, such as the knee receiving strong responses from neighbouring hip/foot joints at adjacent frames. This observation supports our motivation

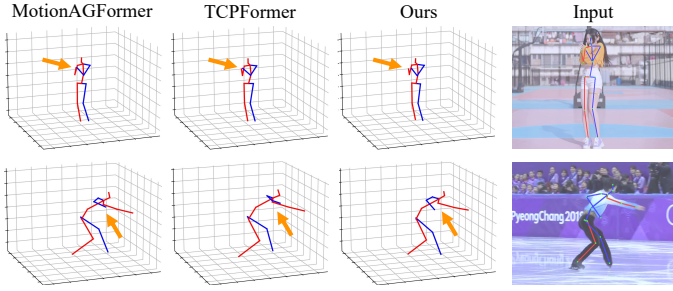


Figure 3: Qualitative comparison against TCPFormer [20] and MotionAGFormer [23] on in-the-wild videos. Orange arrows highlight challenging pose regions.

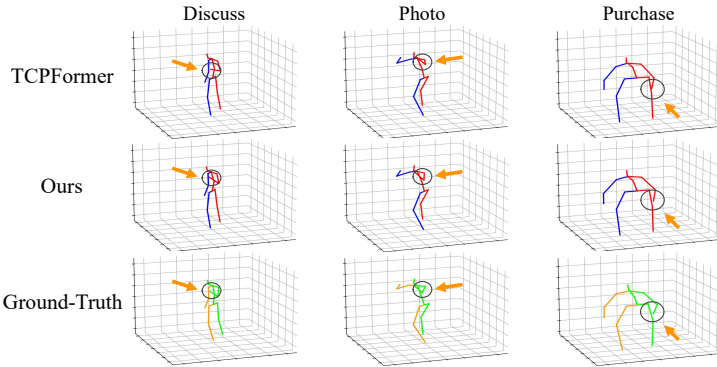


Figure 4: Qualitative comparison against TCPFormer [20] on the Human3.6M dataset. Orange arrows highlight challenging pose regions.

that local pose dynamics are inherently coupled in both spatial and temporal dimensions. By extending the skeleton graph into localised temporal neighbourhoods, HSTG can capture such cross-frame structural dependencies while still preserving anatomical constraints.

4.4.3 Adaptive Fusion Weights Analysis

Figure 5 (b) shows the adaptive fusion weights assigned to HSTG and ADSTG across different layers and actions. The weights are not fixed, but vary with both network depth and motion context. For *SittingDown*, the model assigns relatively higher HSTG weights in deeper layers, suggesting that local coupled spatial-temporal reasoning becomes more important for complex body articulation. In contrast, for *Photo*, the ADSTG weight increases in deeper layers, indicating a stronger reliance on adaptive temporal dependency modelling.

4.4.4 Broader Impact Discussion

Figure 6 shows qualitative zero-shot visualisations beyond human pose estimation using the Human3.6M-trained checkpoint without additional fine-tuning. Although the proposed framework is trained only on human pose data, it can produce visually coherent 3D structures for articulated humanoid robots and bees. These preliminary examples are not intended as

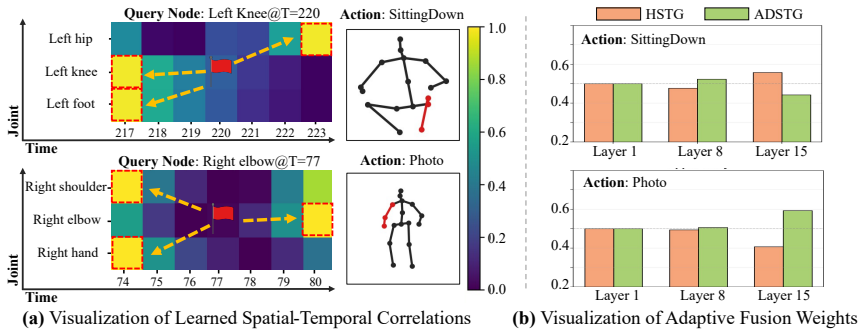


Figure 5: **(a)** Visualisation of learned spatial-temporal correlations in HSTG. Highlighted regions indicate that the query node attends to anatomically related joints across nearby timesteps. **(b)** Visualisation of adaptive fusion weights across different layers and actions.

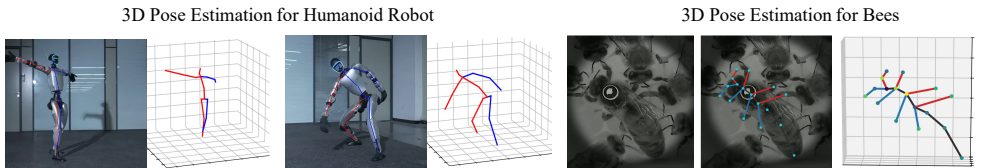


Figure 6: Zero-shot qualitative visualisations on humanoid robots and bees using the Human3.6M-trained checkpoint.

quantitative validation, but rather illustrate the potential of graph-enhanced spatial-temporal reasoning for broader articulated structure modelling.

5 Conclusion

In this paper, we propose HSTGFormer, a graph-enhanced Transformer framework for efficient monocular 3D human pose estimation. Unlike conventional spatial-then-temporal modelling strategies, the proposed framework reformulates spatial-temporal reasoning from a graph perspective through two complementary modules: the Hyper Spatial-Temporal Graph (HSTG) and the Adaptive Dual-Scale Temporal Graph (ADSTG). HSTG extends conventional skeleton graphs into localised temporal neighbourhoods to enable efficient coupled spatial-temporal reasoning, while ADSTG performs adaptive temporal dependency modelling over complementary temporal ranges. In addition, a lightweight node-wise fusion mechanism dynamically integrates the two graph representations according to different motion contexts. Extensive experiments on Human3.6M and MPI-INF-3DHP demonstrate that HSTGFormer achieves strong accuracy while maintaining high computational efficiency and low memory cost. Further ablation studies and qualitative analyses verify the effectiveness of the proposed graph reasoning design. Preliminary zero-shot results on humanoid robots and bees also suggest its potential applicability beyond human pose estimation.

Acknowledgements

This work was supported by EU H2020-FET RoboRoyale project (964492).

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv*, 2016.
- [2] Qiongjie Cui, Huaijiang Sun, Jianfeng Lu, Weiqing Li, Bin Li, Hongwei Yi, and Haofan Wang. Test-time personalizable forecasting of 3d human poses. In *ICCV*, pages 274–283, 2023.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.
- [5] Jia Gong, Lin Geng Foo, Zhipeng Fan, QiuHong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, pages 13041–13051, 2023.
- [6] Nate Hagbi, Oriel Bergig, Jihad El-Sana, and Mark Billinghurst. Shape recognition and pose estimation for mobile augmented reality. *TVCG*, 17(10):1369–1379, 2010.
- [7] Rongtian Huo, Qing Gao, Jing Qi, and Zhaojie Ju. 3d human pose estimation in video for human-computer/robot interaction. In *ICIRA*, pages 176–187, 2023.
- [8] Christian Keilstrup Ingwersen, Christian Møller Mikkelsen, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjørholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *CVPR*, pages 5219–5228, 2023.
- [9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013.
- [10] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [11] Ruochen Li, Stamos Katsigiannis, and Hubert PH Shum. Multiclass-sgc: Sparse graph-based trajectory prediction with agent class embedding. In *ICIP*, pages 2346–2350, 2022.
- [12] Ruochen Li, Stamos Katsigiannis, Tae-Kyun Kim, and Hubert PH Shum. Bp-sgc: Behavioral pseudo-label informed sparse graph convolution network for pedestrian and heterogeneous trajectory prediction. *TNNLS*, 2025.

- [13] Ruochen Li, Tanqiu Qiao, Stamos Katsigiannis, Zhanxing Zhu, and Hubert PH Shum. Unified spatial-temporal edge-enhanced graph networks for pedestrian trajectory prediction. *TCSVT*, 2025.
- [14] Ruochen Li, Ziyi Chang, Junyan Hu, Jiannan Li, Amir Atapour-Abarghouei, and Hubert PH Shum. Art: Adaptive relational transformer for pedestrian trajectory prediction with temporal-aware relations. In *ICHMS*, pages 251–256, 2026.
- [15] Ruochen Li, Shuang Chen, Farshad Arvin, Amir Atapour-Abarghouei, et al. Motion-adaptive multi-scale temporal modelling with skeleton-constrained spatial graphs for efficient 3d human pose estimation. *IJCNN*, 2026.
- [16] Ruochen Li, Zhanxing Zhu, Tanqiu Qiao, and Hubert PH Shum. Vite: Virtual graph trajectory expert router for pedestrian trajectory prediction. In *AAAI*, volume 40, pages 17535–17543, 2026.
- [17] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347, 2014.
- [18] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156, 2022.
- [19] Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Shijian Lu, and Nicu Sebe. H2ot: Hierarchical hourglass tokenizer for efficient video pose transformers. *TPAMI*, 2025.
- [20] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. Tcformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. In *AAAI*, volume 39, pages 5478–5486, 2025.
- [21] Ivan Marisca, Jacob Bamberger, Cesare Alippi, and Michael M Bronstein. Over-squashing in spatiotemporal graph neural networks. In *NeurIPS*, 2025.
- [22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2640–2649, 2017.
- [23] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *WACV*, pages 6920–6930, 2024.
- [24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, pages 506–516, 2017.
- [25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.
- [26] Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *CVPR*, pages 1123–1132, 2024.

- [27] Tanqiu Qiao, Ruochen Li, Frederick WB Li, and Hubert PH Shum. From category to scenery: An end-to-end framework for multi-person human-object interaction recognition in videos. In *ICPR*, pages 262–277, 2024.
- [28] Tanqiu Qiao, Ruochen Li, Frederick WB Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert PH Shum. Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos. *ESWA*, 2025.
- [29] Nathaniel Rossol, Irene Cheng, and Anup Basu. A multisensor technique for gesture recognition through intelligent skeletal pose analysis. *Trans. Hum.-Mach. Syst.*, 46(3): 350–359, 2015.
- [30] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *ECCV*, pages 461–478, 2022.
- [31] Tomohiro Suzuki, Ryota Tanaka, Calvin Yeung, and Keisuke Fujii. Athleticspose: Authentic sports motion dataset on athletic field and evaluation of monocular 3d pose estimation ability, 2025.
- [32] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *CVPR*, pages 4790–4799, 2023.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.
- [34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv*, 2017.
- [35] Ji Yang, Youdong Ma, Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. 3d pose estimation and future motion prediction from 2d images. *PR*, 124:108439, 2022.
- [36] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. In *NeurIPS*, 2023.
- [37] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *ICCV*, pages 8818–8829, 2023.
- [38] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, pages 13222–13232, 2022.
- [39] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *CVPR*, pages 8877–8886, 2023.
- [40] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021.

- [41] Kaili Zheng, Feixiang Lu, Yihao Lv, Liangjun Zhang, Chenyi Guo, and Ji Wu. 3d human pose estimation via non-causal retentive networks. In *ECCV*, pages 111–128, 2024.
- [42] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, pages 15085–15099, 2023.
- [43] Yiran Zhu, Xing Xu, Fumin Shen, Yanli Ji, Lianli Gao, and Heng Tao Shen. Posegtac: Graph transformer encoder-decoder with atrous convolution for 3d human pose estimation. In *IJCAI*, pages 1359–1365, 2021.