Motion-Adaptive Multi-Scale Temporal Modelling with Skeleton-Constrained Spatial Graphs for Efficient 3D Human Pose Estimation

Ruochen Li, Shuang Chen, Wenke E, Farshad Arvin, Amir Atapour-Abarghouei
Department of Computer Science, Durham University, UK
{ruochen.li, shuang.chen, wenke.e, farshad.arvin, amir.atapour-abarghouei}@durham.ac.uk

***Abstract*—Accurate 3D human pose estimation from monocular videos requires effective modelling of complex spatial and temporal dependencies. However, existing methods often face challenges in efficiency and adaptability when modelling spatial and temporal dependencies, particularly under dense attention or fixed modelling schemes. In this work, we propose MASC-Pose, a Motion-Adaptive multi-scale temporal modelling framework with Skeleton-Constrained spatial graphs for efficient 3D human pose estimation. Specifically, it introduces an Adaptive Multi-scale Temporal Modelling (AMTM) module to adaptively capture heterogeneous motion dynamics at different temporal scales, together with a Skeleton-constrained Adaptive GCN (SAGCN) for joint-specific spatial interaction modelling. By jointly enabling adaptive temporal reasoning and efficient spatial aggregation, our method achieves strong accuracy with high computational efficiency. Extensive experiments on Human3.6M and MPI-INF-3DHP datasets demonstrate the effectiveness of our approach.**

***Index Terms*—3D Human Pose Estimation, Graph Neural Network, Spatial-Temporal Learning**

I. Introduction

3D human pose estimation seeks to recover articulated 3D joint coordinates from monocular 2D images or video sequences, and serves as a fundamental component in a wide range of computer vision applications, including human action recognition [1]–[3], human-computer interaction [4]–[6], and virtual reality [7]. Despite its broad applicability, this task remains highly challenging. The complex spatial-temporal dependencies among articulated body joints raise significant challenges, as spatial correlations across limbs and temporal motion patterns are tightly coupled [8], [9]. This coupling makes it difficult to model fine-grained joint interactions and long-range motion dynamics, which is ill-posed and benefits from structured priors that simplify the learning objectives [10].

Recent advances in 3D human pose estimation increasingly adopt Transformer-based architectures to model spatial-temporal dependencies among body joints [11]–[15]. Representative works such as PoseFormer [11] and MixSTE [13] leverage Transformer architectures to model spatial-temporal dependencies, capturing joint-wise correlations and long-range motion dynamics through self-attention mechanisms. As a result, these approaches achieve strong performance through global motion aggregation, but largely rely on vanilla Transformer designs with limited inductive biases. In parallel, motivated by the structured topology of the human skeleton, graph-based methods model joints as nodes and skeletal connections as edges to encode spatial priors. Early work such as ST-GCN [16] demonstrates the effectiveness of Graph Neural Networks (GNNs) in capturing human skeletal structure, and subsequent studies further integrate GNNs with Transformer architectures to enhance spatial interaction modelling [17]–[19]. By explicitly encoding skeletal connectivity, these approaches introduce structural inductive biases that improve spatial reasoning and reduce the reliance on dense attention.

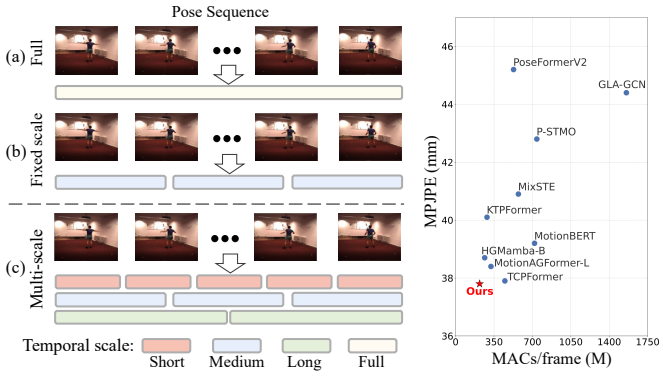

Fig. 1. Comparison of temporal dependency modelling strategies and efficiency–accuracy trade-offs. **Left**: (a) full-sequence, (b) fixed-scale, and (c) the proposed multi-scale temporal modelling. **Right**: Comparison of recent methods on Human3.6M in terms of MPJPE (lower is better) versus MACs/frame, showing that our method achieves state-of-the-art performance with low computational cost.

Despite their strong performance, existing methods still exhibit notable limitations in spatial-temporal learning. For **temporal dependency modelling**, most approaches rely on global temporal attention [11], [17], [18], [20] as shown in Figure 1 (a). While effective in capturing long-range dependencies, applying global temporal attention over long sequences incurs high computational cost and tends to dilute fine-grained temporal saliency due to uniform token aggregation [21]. TCPFormer [15] partially alleviates this issue by capturing fine-grained temporal patterns through fixed-length temporal abstractions (Figure 1(b)); however, the temporal scale remains fixed and dense attention still incurs quadratic complexity and limited adaptability to diverse motion dynamics. For **spatial**

joint interaction modelling, early attention-based methods apply attention over all joints to capture spatial interactions, but stacking attention layers often leads to over-smoothing [22], where joint representations become increasingly similar as depth increases. Recent approaches [17], [18] integrate graph neural networks with skeleton topology. However, their fixed fusion strategies make it difficult to adapt to joint-specific needs in balancing local feature propagation.

In this paper, we propose **MASC-Pose** for 3D human pose estimation. For the **temporal aspect**, MASC-Pose introduces an Adaptive Multi-scale Temporal Modelling (AMTM) module to capture motion dynamics at diverse temporal scales. The key insight is that human motion exhibits heterogeneous temporal patterns, where both short-term dynamics and long-term trends are essential. Accordingly, AMTM processes the input sequence through parallel temporal branches with different scales to model multi-scale motion patterns (Figure 1 (c)). Inspired by the mixture-of-experts (MoE) paradigm [23], [24], we incorporate a learnable fusion mechanism that dynamically weights each temporal scale based on human pose characteristics, while employing lightweight temporal GCNs within each scale to further reduce computational overhead. For the **spatial aspect**, we propose a Skeleton-constrained Adaptive GCN (SAGCN) to model joint interactions. SAGCN introduces learnable balancing weights that adaptively regulate the contributions of self-features and neighbour aggregation at each layer for joint-specific spatial interaction modelling. The right panel of Figure 1 demonstrates the superior performance and efficiency of our approach. *The source code is available on GitHub.* Our contributions are:

- We propose MASC-Pose, a novel spatial-temporal framework for 3D human pose estimation that achieves strong performance with high computational efficiency.
- We introduce an Adaptive Multi-scale Temporal Modelling (AMTM) module that enables efficient and expressive modelling of temporal dependencies at multiple temporal scales (Section III-C), which adaptively balances short-term dynamics and long-term motion dynamics.
- We propose a Skeleton-constrained Adaptive GCN (SAGCN) for efficient and adaptive spatial joint interaction modelling (Section III-B), which enables joint-specific feature aggregation.

II. Related Work

A. Transformer-based Methods

With the success of Transformer architectures [25] in natural language processing, self-attention has become a dominant paradigm for modelling spatial-temporal dependencies in 3D human pose estimation. Early Transformer-based approaches typically adopt factorized spatial-temporal attention, where spatial attention captures joint-wise relationships within each frame and temporal attention models motion dynamics across frames [11]–[14]. To enhance representation capacity, MotionBERT [20] leverages large-scale motion pretraining to learn more robust spatial-temporal representations. More recent studies such as TCPFormer [15] further explore temporal abstraction strategies to capture fine-grained short-term motion details and long-range dependencies jointly. Overall, these works demonstrate the effectiveness of Transformer-based architectures for pose estimation. However, in contrast to these generic temporal modelling schemes, MASC-Pose introduces a motion-adaptive multi-scale temporal module together with skeleton-constrained spatial aggregation to achieve efficient and adaptable spatial-temporal representation learning.

B. Graph-based Methods

Motivated by the structured topology of the human skeleton, graph-based methods represent joints as nodes and skeletal connections as edges to model spatial dependencies. GNN have been widely adopted to capture joint interactions and encode skeletal priors in 3D human pose estimation. Early work such as GLA-GCN [26] combines GNNs with Temporal Convolutional Networks to capture spatial and temporal correlations. Recent approaches integrate GNNs with Transformer architectures [17]–[19], [27], leveraging graph-based spatial modelling to complement attention-based temporal representations. By explicitly modelling joint connectivity, these methods enhance spatial interaction modelling and reduce the computational overhead associated with dense attention mechanisms. Building on this line of work, MASC-Pose retains the skeletal topology prior but makes the spatial message passing more adaptive, while pairing it with motion-aware multi-scale temporal modelling to better handle diverse dynamics under a more efficient spatial-temporal design.

III. Methods

A. Problem Formulation

Given an input 2D human pose sequence $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times J\times 3}$, which contains 2D coordinates and a confidence score of each keypoint. Our objective is to estimate the corresponding 3D pose sequence $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times J\times C_{\text{out}}}$. Here, T denotes the number of frames, J denotes the number of body joints per frame, C_{in} and C_{out} represent the dimensionality of 2D and 3D joint coordinates, respectively. We first project the input 2D poses into high-dimensional feature space as $\mathbf{X}_h \in \mathbb{R}^{T\times J\times D}$ and incorporate positional encoding as $\mathbf{X}_{st} = \Phi\big(\mathbf{X}_h + \mathbf{P}_{\text{pos}}\big)$, where $\mathbf{P}_{\text{pos}} \in \mathbb{R}^{1\times J\times D}$ denotes the joint-level positional encoding, D is the hidden dimension, and $\Phi(\cdot)$ represents the backbone spatial-temporal encoder following [15], [18].

B. Skeleton-constrained Adaptive GCN (SAGCN)

Existing methods for spatial joint modelling commonly rely on self-attention or graph-based formulations. Attention-based approaches compute dense joint interactions with high computational cost and limited use of skeletal priors [11], [15], while graph-based methods explicitly encode skeleton topology but typically adopt fixed aggregation schemes that treat self and neighbour features uniformly [17], [18]. Such uniform aggregation overlooks joint-wise heterogeneity in motion patterns. Motivated by this observation, we propose a Skeleton-constrained Adaptive GCN (SAGCN) for spatial

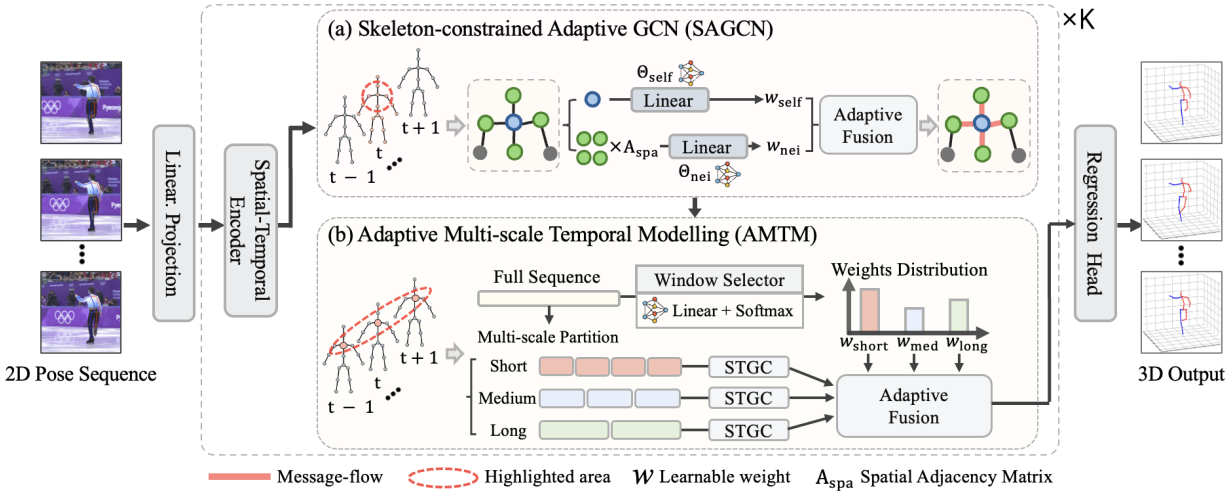

Fig. 2. Overview of the proposed framework. The model integrates (a) a Skeleton-constrained Adaptive GCN (SAGCN) for spatial modelling and (b) an Adaptive Multi-scale Temporal Modelling (AMTM) module for temporal modelling. STGC denotes sparse temporal graph convolution operation.

interaction modelling (Figure 2(a)), which adaptively balances self and neighbour information.

Given input features $\mathbf{X}_h \in \mathbb{R}^{T\times J\times D}$, SAGCN processes each frame independently to model spatial joint interactions by decoupling self-transformation and neighbour aggregation:

$$\mathbf{H} = \sigma\Big(w_{\text{self}} \cdot \mathbf{\Theta}_{\text{self}}(\mathbf{X_h}) + w_{\text{nei}} \cdot \mathbf{\Theta}_{\text{nei}}(\mathbf{A}_{\text{spa}}\mathbf{X_h})\Big), \quad (1)$$

where $\mathbf{\Theta}_{\text{self}}$ and $\mathbf{\Theta}_{\text{nei}}$ are independent learnable linear transformations, $\mathbf{A}_{\text{spa}} \in \mathbb{R}^{J\times J}$ is the normalised skeletal adjacency matrix derived from the skeleton topology, w_{self} and w_{nei} are learnable balancing weights normalised via softmax to control the contribution of self-features and neighbour aggregation, and $\sigma(\cdot)$ denotes the ReLU activation. Unlike standard GCNs that implicitly mix self and neighbour information with fixed contributions, SAGCN allows the network to learn an optimal balance between local joint features and skeletal context.

C. Adaptive Multi-scale Temporal Modelling (AMTM)

For temporal modelling in human pose estimation, most existing methods employ global self-attention [11], [13], [17], [18] or fixed temporal scales [15] to capture temporal correlations. However, human motion exhibits inherently multi-scale characteristics, as rapid movements like hand gestures require short-term modelling, while periodic patterns such as walking benefit from long-term context. Processing all frames with a single temporal scale may fail to capture this diversity and introduce unnecessary computational overhead. Inspired by the routing mechanism of MoE [23], we propose Adaptive Multi-scale Temporal Modelling (AMTM) (Figure 2(b)) that dynamically routes features across multiple temporal scales, implemented via temporal windows of different lengths.

Specifically, AMTM models temporal dependencies using three parallel partitions with different temporal scales such as short-, medium-, and long-term scales. Given the output $\mathbf{H}$ from SAGCN, AMTM employs a lightweight window selector to adaptively estimate the importance of each window:

$$\mathbf{w} = \text{Softmax}\big(\text{MLP}(\text{GAP}(\mathbf{H}))\big), \quad (2)$$

where $\text{GAP}(\cdot)$ denotes global average pooling over temporal dimensions, and $\mathbf{w} = [w_{\text{short}}, w_{\text{med}}, w_{\text{long}}]$ represents the predicted importance weights of each scale.

For each temporal scale, we apply Sparse Temporal Graph Convolution (STGC) [18]. It first constructs a dynamic temporal graph $\mathbf{A}_{\text{temp}}$ by computing pairwise cosine similarities between all timesteps and retaining only the top-k most similar neighbours for each timestep:

$$\mathbf{S} \in \mathbb{R}^{w\times w}, \quad \mathbf{S}(i,j) = \text{cosine}(\mathbf{h}_i, \mathbf{h}_j), \quad (3)$$

$$\mathcal{N}_k(t) = \text{TopK}\big(\mathbf{S}(t,:), k\big), \quad (4)$$

$$\mathbf{A}_{\text{temp}}(t,\tau) = \begin{cases} 1, & \tau \in \mathcal{N}_k(t), \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

The sparse adjacency matrix is normalised and used to perform temporal graph convolution:

$$\text{STGC}(\mathbf{A}_{\text{temp}}, \mathbf{H}_w) = \sigma\Big(\text{BN}\big(\tilde{\mathbf{A}}_{\text{temp}}\mathbf{\Theta}_{\text{nei}}(\mathbf{H}_w) + \mathbf{\Theta}_{\text{self}}(\mathbf{H}_w)\big)\Big), \quad (6)$$

where $\tilde{\mathbf{A}}_{\text{temp}}$ is the normalised adjacency matrix, $\mathbf{\Theta}_{\text{nei}}$ and $\mathbf{\Theta}_{\text{self}}$ are learnable transformations for neighbour and self features, and $\sigma(\cdot)$ is the ReLU activation. The output of each scale is represented as:

$$\mathbf{H}_{\text{short}} = \text{STGC}_{\text{short}}(\mathbf{A}^{\text{short}}_{\text{temp}}, \text{Partition}_{\text{short}}(\mathbf{H})), \quad (7)$$

$$\mathbf{H}_{\text{med}} = \text{STGC}_{\text{med}}(\mathbf{A}^{\text{med}}_{\text{temp}}, \text{Partition}_{\text{med}}(\mathbf{H})), \quad (8)$$

$$\mathbf{H}_{\text{long}} = \text{STGC}_{\text{long}}(\mathbf{A}^{\text{long}}_{\text{temp}}, \text{Partition}_{\text{long}}(\mathbf{H})), \quad (9)$$

where $\text{Partition}_{\text{scale}}(\cdot)$ reshapes the temporal dimension into non-overlapping windows to allow for efficient and focused temporal modelling at each scale. Finally, we adaptively aggregate the multi-scale outputs:

$$\mathbf{X}_{\text{out}} = w_{\text{short}} \cdot \mathbf{H}_{\text{short}} + w_{\text{med}} \cdot \mathbf{H}_{\text{med}} + w_{\text{long}} \cdot \mathbf{H}_{\text{long}}, \quad (10)$$

TABLE I

QUANTITATIVE COMPARISONS ON THE HUMAN3.6M DATASET. CE INDICATES WHETHER THE MODEL ESTIMATES THE CENTER FRAME ONLY. T DENOTES THE NUMBER OF INPUT FRAMES. MACS/FRAMES REPRESENTS THE NUMBER OF MULTIPLY-ACCUMULATE OPERATIONS PER OUTPUT FRAME. MPJPE† INDICATES THAT GROUND-TRUTH 2D POSES ARE USED AS INPUT. THE BEST RESULTS ARE SHOWN IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Venue	CE	T	Parameter	MACs	MACs/frames	MPJPE ↓	P-MPJPE ↓	MPJPE† ↓
MixSTE [13]	CVPR'22	✗	243	33.6M	139.0G	572M	40.9	32.6	21.6
P-STMO [28]	ECCV'22	✓	243	6.2M	0.7G	740M	42.8	34.4	29.3
PoseFormerV2 [12]	CVPR'23	✓	243	14.3M	0.5G	528M	45.2	35.6	–
GLA-GCN [26]	ICCV'23	✓	243	1.3M	1.5G	1556M	44.4	34.8	21.0
MotionBERT [20]	ICCV'23	✗	243	42.3M	174.8G	719M	39.2	32.9	17.8
KTPFormer [17]	CVPR'24	✗	243	33.7M	69.5G	286M	40.1	31.9	19.0
MotionAGFormer-L [18]	CVPR'24	✗	243	19.0M	78.3G	322M	38.4	32.5	17.3
HGMamba-B [29]	IJCNN'25	✗	243	14.2M	64.5G	265M	38.7	32.9	**13.2**
H2OT + MotionAGFormer [30]	TPAMI'25	✗	243	11.7M	38.9G	-	38.5	-	-
TCPFormer [15]	AAAI'25	✗	243	35.1M	109.2G	449M	<u>37.9</u>	**31.7**	<u>15.5</u>
Ours	–	✗	243	13.0M	53.4G	219M	**37.8**	<u>31.8</u>	16.0

where $\mathbf{X}_{\text{out}}$ represents the aggregated feature representation. Finally, following [15], [18], a softmax-based adaptive fusion is used to combine $\mathbf{x}_{\text{out}}$ and $\mathbf{x}_{\text{st}}$, producing the final embedding $\mathbf{X}_{\text{final}}$, which is fed into a linear regression head to estimate the final 3D pose sequence. A progressive layer fusion is adopted to aggregate features across layers to preserve multi-level representations and improve gradient flow.

D. Loss Function

Following prior work [15], our model is trained end to end. The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_m + \lambda_s \mathcal{L}_s + \lambda_v \mathcal{L}_v + \lambda_d \mathcal{L}_d, \tag{11}$$

where $\mathcal{L}_m$ denotes MPJPE, $\mathcal{L}_s$ represents N-MPJPE, $\mathcal{L}_v$ is the velocity loss, and $\mathcal{L}_d$ represents temporal consistency loss. We set λ_s to 0.5, λ_v to 20, and λ_d to 0.5, respectively.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate our model on two widely used 3D human pose estimation benchmarks, Human3.6M [31] and MPI-INF-3DHP [32]. Human3.6M is a standard indoor dataset with multi-view recordings of multiple subjects performing common actions, while MPI-INF-3DHP covers both indoor and outdoor scenes with diverse activities. Following standard evaluation protocols [15], [18], we report MPJPE and Procrustes-aligned MPJPE (P-MPJPE) on Human3.6M. For MPI-INF-3DHP, ground-truth 2D poses are used as input, and performance is evaluated using MPJPE, Percentage of Correct Keypoints (PCK), and Area Under the Curve (AUC).

B. Implementation Details

Our model is implemented in PyTorch and trained on a single NVIDIA H200 GPU. We use 16 encoder layers with a feature dimension of 128. The input sequence length is 243 for Human3.6M and 81 for MPI-INF-3DHP, with temporal scales set to $\{9, 27, 81\}$ and $\{3, 9, 27\}$, respectively. Data preprocessing and evaluation protocols follow [15]. Training is performed using AdamW with a learning rate of 5×10^{-4} for 80 epochs and a batch size of 6.

TABLE II

QUANTITATIVE COMPARISON ON MPI-INF-3DHP DATASET. T DENOTES THE NUMBER OF INPUT FRAMES. PCK AND AUC ARE REPORTED IN PERCENTAGE (%). THE BEST RESULTS ARE SHOWN IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	Venue	T	PCK↑	AUC↑	MPJPE↓
MixSTE [13]	CVPR'22	27	94.4	66.5	54.9
P-STMO [28]	ECCV'22	81	97.9	75.8	32.2
PoseFormerV2 [12]	CVPR'23	81	97.9	78.8	27.8
GLA-GCN [26]	ICCV'23	81	98.5	79.1	27.7
KTPFormer [17]	CVPR'24	81	98.9	85.9	16.7
MotionAGFormer-L [18]	WACV'24	81	98.2	85.3	16.2
HGMamba-B [29]	IJCNN'25	81	98.7	<u>87.9</u>	**14.3**
H2OT + MotionAGFormer [30]	TPAMI'25	81	**99.1**	85.2	18.0
TCPFormer [15]	AAAI'25	81	<u>99.0</u>	87.7	<u>15.0</u>
Ours	–	81	**99.1**	**88.2**	15.5

C. Quantitative Comparison with State-of-the-art Methods

1) Results on Human3.6M dataset: Table I reports comparisons with state-of-the-art methods on the Human3.6M dataset, evaluating both computational cost and prediction accuracy in terms of MPJPE, P-MPJPE and MPJPE†. Our method achieves the best MPJPE among all compared approaches, while maintaining competitive performance on P-MPJPE and MPJPE†. Notably, compared with recent transformer-based methods such as KTPFormer and TCPFormer, our approach attains comparable or superior accuracy with substantially fewer parameters and lower computational cost.

2) Results on MPI-INF-3DHP dataset: Table II reports quantitative comparisons with state-of-the-art methods on the MPI-INF-3DHP dataset using PCK, AUC, and MPJPE. Our method achieves competitive performance across all metrics, attaining the highest AUC and matching the best PCK score with a low MPJPE. Compared with TCPFormer, our approach delivers comparable or improved accuracy with a more efficient spatial-temporal modelling strategy, indicating strong generalization ability to challenging scenarios.

TABLE III
ABLATION STUDY OF DIFFERENT MODEL VARIANTS ON HUMAN3.6M.

Variant	MPJPE ↓
Baseline (w/o AMTM & SAGCN)	41.1
Only AMTM	38.5
Only SAGCN	39.9
AMTM + SAGCN (w/o adaptive aggregation)	38.3
AMTM (w/o multi-scale fusion) + SAGCN	38.6
AMTM + SAGCN (**Ours**)	**37.8**

TABLE IV
ABLATION ON TEMPORAL SCALE ON HUMAN3.6M WITH $T = 243$.

Temporal scales	MPJPE ↓
[3, 9, 27]	38.7
[27, 81, 243]	38.6
[9, 27, 81] (**Ours**)	**37.8**

TABLE V
ABLATION ON SAGCN HOP NUMBER.

Hop number (K)	MPJPE ↓
$K = 4$	39.0
$K = 3$	38.5
$K = 2$	38.2
$K = 1$ (**Ours**)	**37.8**

D. Ablation Study and Analysis

We conduct a series of ablation study on Human3.6M [31] to evaluate the effectiveness of main modules of our method.

1) Impact of main component: Table III examines the contribution of individual components in our framework on Human3.6M. Relative to the baseline, incorporating AMTM alone leads to a substantial reduction in MPJPE to 38.5, which is notably larger than the improvement achieved by using SAGCN alone, suggesting that temporal modelling plays a more prominent role in this task. Combining both components yields further performance gains, while removing the adaptive mechanisms consistently results in degraded accuracy. specifically, replacing the adaptive multi-scale aggregation operation in AMTM with fixed weights (i.e. $1/3$ for each temporal scale) leads to a performance degradation to 38.6, indicating the importance of adaptive scale weighting mechanism. The full model achieves the best overall performance, demonstrating the complementary contributions of AMTM and SAGCN.

2) Effect of Temporal Scale Configuration: Table IV presents an ablation study on temporal scale configurations on the Human3.6M dataset with $T = 243$. Using small temporal scales $[3, 9, 27]$ limits the model's ability to capture long-term motion dependencies, while overly larger scales $[27, 81, 243]$ tend to dilute fine-grained temporal dynamics. The proposed configuration achieves the best performance, indicating that a balanced combination and design of different scales can be crucial for improved performance.

3) Impact of SAGCN Hop Configuration: As shown in Table V, we study the effect of the hop number K in SAGCN, which controls the spatial propagation range, with $K = 1$ aggregating information from each joint and its directly connected neighbours defined by the skeleton topology. The results suggest that larger spatial receptive fields do not necessarily improve performance, likely due to over-smoothing effects [22] caused by increasing the number of GNN layers.

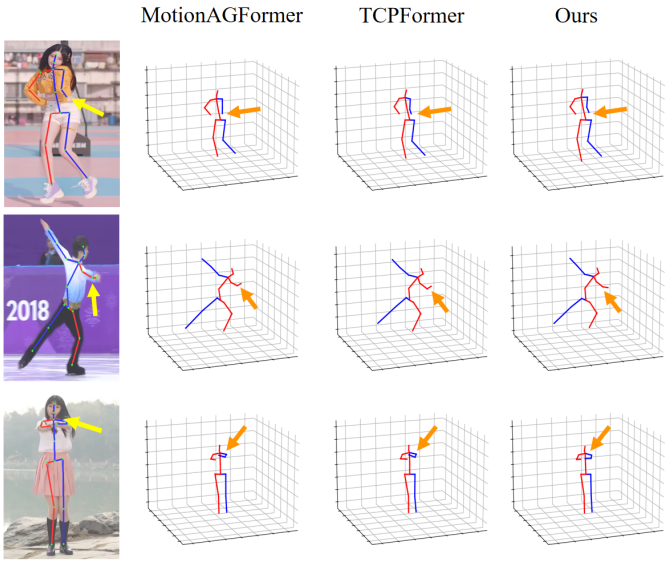

Fig. 3. Qualitative comparisons of our method with MotionAGFormer and TCPFormer on in-the-wild videos. We highlight inaccurate or ambiguous 2D detections with light-yellow arrows and indicate the corresponding deviations in the reconstructed 3D poses using orange arrows.

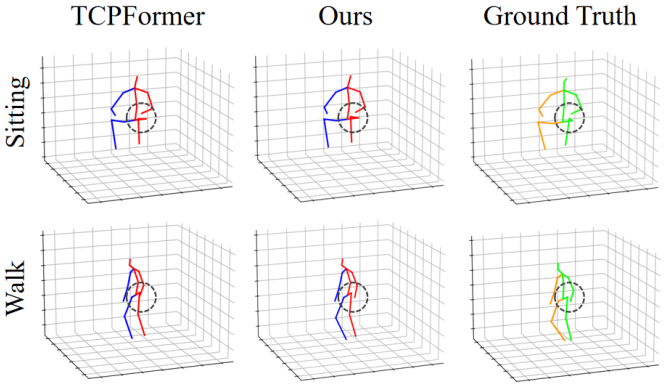

Fig. 4. Qualitative comparisons between our method and TCPFormer on the Human3.6M dataset for the *Sitting* and *Walk* actions. Black dashed circles indicate highlighted regions.

E. Qualitative Analysis

Figure 3 shows qualitative comparisons with MotionAGFormer [18] and TCPFormer [15] on in-the-wild videos. Compared methods exhibit noticeable 3D pose deviations under inaccurate or ambiguous 2D detections, whereas our method produces more stable and anatomically plausible pose estimation results. Figure 4 shows comparisons between our method and TCPFormer on the Human3.6M dataset for the *Sitting* and *Walk* actions. Our method shows 3D pose estimations that are closer to the ground truth compared to TCPFormer.

As shown in Figure 5, we analyse the learned temporal scale weights across different body parts for *Walk* and *SittingDown* actions on Human3.6M with $T = 243$. The two actions exhibit distinct temporal characteristics: *Walk* involves periodic and repetitive motion patterns, whereas *SittingDown* is characterised by more abrupt and transitional dynamics.

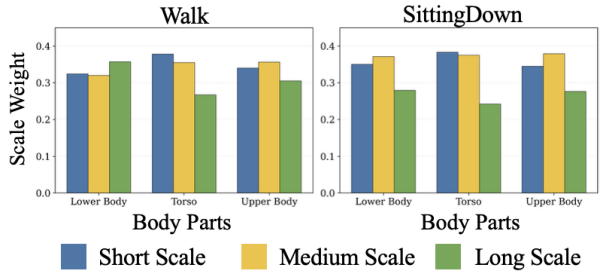

Fig. 5. Visualisation of the scale weight distribution for the *Walk* and *SittingDown* actions on the Human3.6M dataset (243 frames). Body joints are grouped into **lower body** (hip, knee, ankle), **torso** (root, spine, thorax, neck, head), and **upper body** (shoulder, elbow, wrist). Bars represent the average selection weights assigned to three temporal scales: short (9 frames), medium (27 frames), and long (81 frames).

Accordingly, for *Walk*, the lower body assigns higher weights to long temporal scales to capture periodic gait motions, while the torso emphasises short scales for rapid upper-body dynamics. In contrast, for *SittingDown*, all body parts favour short and medium temporal scales, highlighting the importance of intermediate temporal context during posture transitions. These visualisations demonstrate that our model effectively adapts to diverse motion scenarios and enhances the interpretability of the temporal representations.

V. Conclusion

We propose MASC-Pose, an efficient spatial-temporal framework for 3D human pose estimation that addresses the limitations of global or fixed-scale temporal modelling while preserving strong accuracy. By adaptively learning both multi-scale temporal correlations and skeleton-constrained spatial interactions, the proposed method achieves a favorable balance between estimation accuracy and computational efficiency. Extensive experiments demonstrate that our method achieves strong performance and robustness under diverse motion patterns. Future work will explore adaptive temporal partitioning with overlapping or variable-length scales for modelling non-stationary motions, as well as broader generalisation settings such as cross-dataset transfer and in-the-wild evaluation.

References

[1] C. Zhang, T. Yang, J. Weng, M. Cao, J. Wang, and Y. Zou, "Unsupervised pre-training for temporal action localization tasks," in *CVPR*, 2022, pp. 14 011–14 021.

[2] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[3] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *CVPR*, 2018, pp. 1159–1168.

[4] N. Rossol, I. Cheng, and A. Basu, "A multisensor technique for gesture recognition through intelligent skeletal pose analysis," *Trans. Hum.-Mach. Syst.*, vol. 46, no. 3, pp. 350–359, 2015.

[5] R. Huo, Q. Gao, J. Qi, and Z. Ju, "3d human pose estimation in video for human-computer/robot interaction," in *ICIRA*, 2023, pp. 176–187.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011, pp. 1297–1304.

[7] N. Hagbi, O. Bergig, J. El-Sana, and M. Billinghurst, "Shape recognition and pose estimation for mobile augmented reality," *TVCG*, vol. 17, no. 10, pp. 1369–1379, 2010.

[8] J. Lange and M. Lappe, "The role of spatial and temporal information in biological motion perception," *Advances in Cognitive Psychology*, vol. 3, no. 4, p. 419, 2008.

[9] R. Fredericksen, F. Verstraten, and W. Van de Grind, "Spatio-temporal characteristics of human motion perception," *Vision research*, vol. 33, no. 9, pp. 1193–1205, 1993.

[10] A. Atapour-Abarghouei and T. P. Breckon, "Monocular segment-wise depth: Monocular depth estimation based on a semantic segmentation prior," in *ICIP*, 2019, pp. 4295–4299.

[11] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *ICCV*, 2021, pp. 11 656–11 665.

[12] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *CVPR*, 2023, pp. 8877–8886.

[13] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video," in *CVPR*, 2022, pp. 13 222–13 232.

[14] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, "3d human pose estimation with spatio-temporal criss-cross attention," in *CVPR*, 2023, pp. 4790–4799.

[15] J. Liu, M. Liu, H. Liu, and W. Li, "Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation," in *AAAI*, vol. 39, no. 5, 2025, pp. 5478–5486.

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, vol. 32, no. 1, 2018.

[17] J. Peng, Y. Zhou, and P. Mok, "Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation," in *CVPR*, 2024, pp. 1123–1132.

[18] S. Mehraban, V. Adeli, and B. Taati, "Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network," in *WACV*, 2024, pp. 6920–6930.

[19] X. Pan, G. Li, N. Zhang, and J. Li, "3d human pose estimation based on a hybrid approach of transformer and gcn-former," *JVCIR*, p. 104696, 2025.

[20] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *ICCV*, 2023, pp. 15 085–15 099.

[21] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," in *ICML*, 2021, pp. 2793–2803.

[22] X. Wu, A. Ajorlou, Z. Wu, and A. Jadbabaie, "Demystifying over-smoothing in attention-based graph neural networks," *NeurIPS*, vol. 36, pp. 35 084–35 106, 2023.

[23] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *JMLR*, vol. 23, no. 120, pp. 1–39, 2022.

[24] R. Li, Z. Zhu, T. Qiao, and H. P. Shum, "Vite: Virtual graph trajectory expert router for pedestrian trajectory prediction," in *AAAI*, vol. 40, no. 21, 2026, pp. 17 535–17 543.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[26] B. X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, and C. W. Chen, "Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video," in *ICCV*, 2023, pp. 8818–8829.

[27] R. Li, T. Qiao, S. Katsigiannis, Z. Zhu, and H. P. Shum, "Unified spatial-temporal edge-enhanced graph networks for pedestrian trajectory prediction," *TCSVT*, 2025.

[28] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao, "P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation," in *ECCV*, 2022, pp. 461–478.

[29] H. Cui and T. Hayama, "Hgmamba: Enhancing 3d human pose estimation with a hypergcn-mamba network," in *IJCNN*, 2025, pp. 1–8.

[30] W. Li, M. Liu, H. Liu, P. Wang, S. Lu, and N. Sebe, "H_2ot: Hierarchical hourglass tokenizer for efficient video pose transformers," *TPAMI*, 2025.

[31] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2013.

[32] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *3DV*, 2017, pp. 506–516.