A Multi-Granular Joint Tracing Transformer for Video-Based 3D Human Pose Estimation

Yingying Hou[1], Zhenhua Huang[1*], Wentao Zhu[2]

[1]Anhui University, Hefei, 230601, Anhui, China.
[2]Amazon Research, Seattle, 98101, Washington, United States.

*Corresponding author(s). E-mail(s): zhhuangscut@gmail.com;
Contributing authors: hdyyzdn@gmail.com; wentaozhu91@gmail.com;

Abstract

Human pose estimation from monocular images captured by motion capture cameras is a crucial task with a wide range of downstream applications, e.g., action recognition, motion transfer, and movie making. However, previous methods have not effectively addressed the depth blur problem while considering the temporal correlation of individual and multiple body joints together. We address the issue by simultaneously exploiting the temporal information at both single-joint and multiple-joint granularities. Inspired by the observation that different body joints have different moving trajectories and can be correlated with others, we proposed an approach called the Multi-granularity jOint Tracing Transformer (MOTT). MOTT consists of two main components: (1) a spatial transformer that encodes each frame to obtain spatial embeddings of all joints, and (2) a multi-granularity temporal transformer that includes both a holistic temporal transformer to handle the temporal correlation between all joints in consecutive frames and a joint tracing temporal transformer to process the temporal embedding of each particular joint. The outputs of the two branches are fused to produce accurate 3D human poses. Extensive experiments on Human3.6M and MPI-INF-3DHP datasets demonstrate that MOTT effectively encodes the spatial and temporal dependencies between body joints and outperforms previous methods in terms of mean per joint position error.

Keywords: 3D human pose estimation, Joint-Tracing Transformer, Temporal dependencies, Spatial relationship

1 Introduction

Recently, monocular 3D human pose estimation from motion capture cameras has attracted considerable attention in the computer vision community and Internet of Things as it is crucial to various computer vision and graphics applications, *e.g.*, human pose tracking, pose-guided image synthesis, AR/VR applications, and style transformation [1]. The objective of human pose estimation is to automatically locate the 3D positions of human body parts, *e.g.*, neck, and head, from an image or a video of human motion captured by sensors and cameras. However, since a single image can appear as a 2D projection of multiple 3D pose coordinates onto the same 2D plane pose coordinate (known as the depth blur problem), recent advances [2–11] have attempted utilize 2D sequences around a single frame to assist the prediction from a 2D image to a 3D position. In this way, the temporal and spatial information of multiple frames is leveraged to address the problem.

Although existing methods [8, 10] have made progress in addressing this problem from either a temporal or spatial perspective, the temporal correlation of joints has not been fully and effectively exploited in Transformer-based video 3D pose estimation methods. Poseformer [8] proposes a Transformer-based model that estimates the body pose of a center frame from adjacent frames (Fig. 1 (a)). It employs a temporal Transformer to model the global features of all joints across multiple frames altogether. However, different body joints (*e.g.*, elbow, foot, etc) usually have different moving trajectories, and tracking all human joints simultaneously leads to limited exploitation of spatial-temporal correlations. To further exploit these correlations, MixSTE [12] separates the temporal motion of each body joint over a long sequence (Fig. 1 (b)). However, MixSTE discards the dependencies of different body joints, which neglects the fact that the movements of adjacent joints affect each other.

To utilize spatial-temporal correlations of each joint, we propose a novel paradigm that works along both lines, *i.e.*, capturing the temporal correlation of each joint and all joints over multiple frames. Specifically, our proposed paradigm aims to trace the motion of human body joints over multiple frames at the granularity of individual joints and all joints. To achieve this, we introduce a novel Multi-granularity jOint Tracing Transformer (MOTT). As shown in Fig. 1 (c), we first use a spatial Transformer to encode the spatial correlation of body joints in each frame. To fully capture the temporal relationship of human body joints, we design a novel Multi-Granularity Temporal Transformer module (MGTT) to model complicated dependencies of joints across multiple frames. In particular, we model the holistic-granularity dependencies of all joints across various frames via a holistic temporal Transformer, which leverages features of multiple frames. Moreover, to model the local-granularity temporal correlation of an individual joint, which often has individual motions, we propose a joint-tracing temporal Transformer that processes features of each joint across multiple frames in a joint-wise manner. To combine temporal information of joints at both granularities, we use a regression head to decode the temporal features for 3D joint positions.

We demonstrate the effectiveness of MOTT by extensive quantitative and qualitative experiments on widely adopted Human3.6M and MPI-INF-3DHP datasets. Moreover, we evaluate the performance of our method in real-world scenarios by collecting 30 in-the-wild videos. The experiments show that the proposed method, which better encodes the spatial and temporal relationship

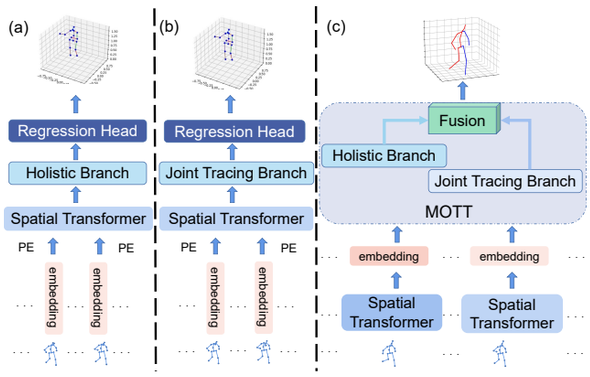

Fig. 1 Compared to previous methods that use the temporality of all human body joints (a) or each joint (b) to model temporal correlation features, our proposed MOTT (c) employs a novel Multi-Granularity Joint Tracing Transformer architecture that captures both holistic and local dependencies of human body joints.

between body joints, accurately estimates human body joints on these large-scale datasets.

The main contributions are as follows:

- We propose a paradigm that leverages the holistic motion of related human body joints and the individual motion of each joint to comprehensively model human pose. This is achieved by learning multi-granularity temporal correlation, which tackles both the holistic temporal relationship and the joint-wise temporal dependencies of human body joints.
- We introduce a multi-granularity temporal Transformer, which includes a holistic temporal Transformer that models frame-wise human joint features, and a joint-tracing temporal Transformer that captures joint-wise human joint features.
- Extensive experiments and comparisons on different datasets demonstrate that our method can effectively estimate 3D human pose from a 2D observation sequence accurately.

2 Related Work

2.1 Monocular Pose Estimation.

Recent advances in monocular pose estimation [13–19] can be primarily divided into end-to-end and 2D-3D lifting approaches.

2D-3D lifting-based methods: Martinez et al. [13] proposed a simple yet very effective fully connected neural network model to predict 3D key point coordinates, which verified the accuracy of the 2D joint one of the sources of an error rate of pose estimation task. Based on the recurrent neural network (RNN), Fang et al. [14] used kinematics prior, symmetry prior, and coordination before predicting human joint coordinates. Panda et al. [20] exploited the deep ambiguity problem inherent in the task and proposed a method for the emergence of multiple feasibility hypotheses during 2D-3D regression.

End-to-end-based methods: Pavlakos et al. [21] adopted an end-to-end method, using the convolution module to output the heatmap of each joint point according to the input RGB image. Sun et al. [22] introduced a novel approach to structure-aware pose regression, which utilizes a reparameterized pose representation that is based on bones instead of joints. Pavlakos et al. [23] proposed to use a weaker supervision signal provided by the ordinal depths of human joints. Through research, it is found that, as the performance of the 2D pose estimation algorithm becomes more powerful, the method [12] of generating 3D pose coordinates using an advanced 2D pose generator is better than the end-to-end method. Liu et al. [19] proposed the absorption graph to focus on specific spatio-temporal relationships in point clouds and voxels. The Absorption Graph Convolutional Network (AGCN) utilizes Graph Convolutional Networks (GCNs) to learn accurate 3D pose estimation. Our method can be primarily considered a 2D-3D lifting-based method.

2.2 Video Pose Estimation.

Video pose estimation provides lateral temporal information and improves the accuracy of pose estimation by using the spatial structure features of the human body in the front and rear frames. Wang et al. [5] proposed a UGCN model, which uses a graph convolutional network (GCN) to capture short-range and long-range motion information, And the smooth line constraint loss function is added to the temporal sequence prediction. Many methods combine the Transformer and pose estimation task to improve the capture ability of the model for long video input frames, Zhao et al. [24] introduced a GraFormer model combining the Transformer model with graph convolution, which fuses the information of all joints to model the topological structure of the human body. Inspired by MAE [25], San et al. [11] designed a self-supervised pre-training P-STMO, which randomly masked the joints of input sequences from both spatial and temporal perspectives. Ma et al. [26] model the skeleton structure of human beings from the perspective of time and space to learn local and global characteristics. They all address the problem of deep blur from the perspectives of both time and space, without tracking the temporal information of joint nodes. Inspired by this, we designed a dual-temporal module to learn the temporal relationship between frames and the same joint sequence, thus enhancing the representation capability of human structural features.

2.3 Dual-Temporal Pose Estimation.

Modeling from the perspective of the spatial domain or temporal-spatial domain is called single-temporal, and modeling temporal information from local and global, which is called dual-temporal. Hossain et al. [27] introduced a recurrent neural network using a Long Short-Term Memory (LSTM) unit with shortcut connections to exploit temporal information from sequences of human pose. Inspired by ViT [28], Li et al. [10] proposed the Strided Transformer, which continuously performs a hierarchical transformation on input sequences and strided convolutions. However, the extraction of the human body's spatial topology information and the learning of time series are not effective. Dabral et al. [29] exploited the spatial-temporal relationships and constraints, e.g., bone-length constraint and left-right symmetry constraint, to improve 3D HPE performance from sequential frames. Different from the above method, our MOTT employs a spatial Transformer module to learn the structural features of a single frame independently, constructs a parallel module to handle local and global joint modeling, and superimposes and fuses double-timing features to improve the prediction accuracy of the center frame.

2.4 Transformer-Based Pose Estimation.

Previous works use RNN and its internal memory to process input sequences of arbitrary time series. Fang et al. [14] established a hierarchical structure of RNNs for generating final high-level 3D attitude grammars for coding rational 3D human estimation. Convolutional neural networks (CNN) also show the ability of multi-scale feature extraction. Pavllo et al. [3] proposed a spatial-temporal dilated convolution model. Compared with RNN, it has higher accuracy, is simpler, and more effective, which has advantages in computational complexity and the number of parameters. Moreover, the model

has a stronger ability to capture long-term information. Inspired by recent developments in vision transformers, Zheng et al. [8] used a Transformer encoder instead of a convolution module to design a spatial-temporal Transformer structure to predict 3D joint coordinates. Ma et al. [30] and Shuai et al. [31] used Transformers to extract multi-view features. Specifically, Ma et al. [30] designed a unified Transformer architecture to fuse cues from both current views and neighboring views. Shuai et al. [31] proposed an MTF-Transformer to adaptively handle varying view numbers and video length without camera calibration in 3D Human Pose Estimation (HPE).

3 METHOD

MOTT consists of two key components: a Spatial Transformer Encoder for learning the spatial structure relationship of body joints in a single frame, and a Multi-Granularity Temporal Transformer for exploiting the temporal correlation of body joints over multiple frames at two different granularities. The overall framework of the proposed method is illustrated in Fig. 2. At the cross-joint granularity, the Holistic Temporal Transformer (HTT) seeks to learn the temporal relationship between related joints across multiple frames. At the single-joint granularity, the Joint Tracing Temporal Transformer (JTTT) aims to exploit the temporal correlation of each joint over different frames.

3.1 Spatial Transformer Encoder

To prepare spatial information for modeling the temporal correlation of body joints at multiple granularities, we deploy a spatial transformer encoder (STE) to learn the structural correlations of all joints in every single frame separately. Specifically, STE encodes the pose coordinates of all skeletal points in the input frame into high dimensions and learns the spatial structure relationship of human body joints. Given a series of two-dimensional joint coordinates, $F = \{f_1, f_2, .., f_J\}$, where J represents the number of joint points, and each joint point contains two dimensions x and y, we increase the number of coordinate features to 32 dimensions through a fully connected layer to improve the representative ability. The STE then learns the dependencies and correlations of body joints as:

$$F_1 = \text{GELU}(w_1 x + b_1) \tag{1}$$

$$F_2 = \text{Dropout}(w_2 F_1 + b_2) \tag{2}$$

Where GELU is the activation function, w_1 represents the characteristics of each joint point of input.

3.2 Multi-Granularity Temporal Transformer

The multi-granularity temporal transformer is designed based on two observations: (1) in a video sequence, different body joints usually have different moving trajectories. (2) the motion of a joint has an impact on adjacent or related joints. Following the observation, we propose to handle the temporal correlation of human body joints at different granularity, *i.e.*, modeling temporal dependencies of each *single joints separately* and *multiple joints altogether* across multiple frames. To this end, we propose a holistic temporal transformer (HTT) and a joint tracing temporal transformer (JTTT).

Holistic Temporal Transformer. HTT uses one-dimensional convolution to learn the temporal relationship between frames of a certain length, which reduces the dimension of the joint feature $C \in R^{J\times 32}$ generated by STE to dim = 256. The transformer uses multi-head attention to model the frame sequence, and a stridden convolution is added to the module to reduce the redundancy of input frames. For training different numbers of input frames, the stridden Convolution size is set differently. The residual layer uses max-pooling to retain the main feature information, and HTT adopts the training method of multiple inputs and a single output to collect 3D attitude information in the center frame comprehensively.

Joint Tracing Temporal Transformer. The JTTT encodes the motion relationship between joints by using a double-branch parallel method to supplement features and independent learning does not affect each other. A series of video frames $L = \{l_1, l_2, .., l_N\}$ are input into the STE module. Based on the joint feature information $T \in R^{J\times N\times 32}$ generated by STE, we compress the dimensions by adjusting the joint coding dimension to 256 dimensions to obtain frame vector features E. We then use a self-attention mechanism to deal with the dependence of arbitrary length sequences and capture remote information to learn joint motion features. Finally, the JTTT model is used to obtain the output joint characteristic information $I \in R^{1\times J\times 256}$.

JTTT and HTT complementarily learn long-term and short-term temporal characteristics in parallel. Given the joint information $C \in R^{N\times 32}$ of all frames learned by STE, where N is the number of input video frames, an aggregation vector function is utilized to aggregate all feature information of the same joint point. Since the aggregated feature dimension reaches 10000 levels, resulting in high computational complexity, so by using one-dimensional convolution to compress the feature dimension to 256, and the compressed feature dimension $T \in R^{N\times 256}$. This approach allows JTTT to independently learn the temporal characteristics of each key point of the central frame from a higher dimension, while enabling HTT to learn sequence information of the entire joints with highly fused features, capturing more implicit relationships. As a result, the 3D feature information of the central frame becomes more abundant. The formula is defined as:

$$Z_\mu = MSA_1(LN(Z_\ell)) + Z_\ell \tag{3}$$

$$Y_\partial = FFN_1(Z_\mu) \tag{4}$$

$$Z_\emptyset = MSA_2(LN(Z_\ell)) + Z_\ell \tag{5}$$

$$Y_\beta = FFN_2(Z_\emptyset) \tag{6}$$

$$A = FFN_3(Concat(Y_\partial, Y_\beta)) \tag{7}$$

In Equation 3, Z_ℓ represents the spatial posture feature generated by STE for a single frame, Z_μ denotes the sequence frame feature vector generated by the attention mechanism in HTT, and in Equation 5, similarly $Z_\emptyset$ represents the node temporal feature vector generated by the attention mechanism network in JTTT.

FFN_1 and FFN_2 represent feedforward neural networks with different numbers of convolutional kernels. To better fuse the features, concatenate is used to join them together, and the FFN_3 network is used to aggregate and generate the posture features of the target frame.

3.3 Overall Loss Function

To train our proposed multi-granularity transformer for human pose estimation, we use Mean Per Joint Position Error (MPJPE) as our overall loss function to minimize the error between all predicted 3D joint coordinates and ground truth attitude coordinates:

$$\mathcal{L} = \frac{1}{J}\sum_{k=1}^{J} \|Y_k - \hat{Y}_k\|_2, \tag{8}$$

Where Y_k represents the 3D joint coordinates of the k_{th}

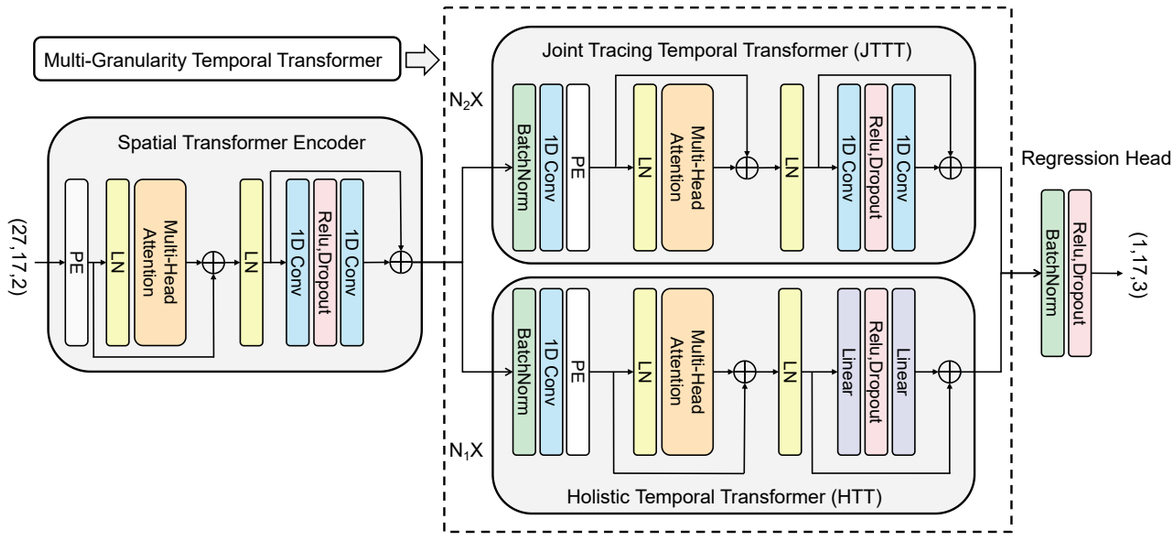

Fig. 2 The overall pipeline of MOTT. MOTT takes a 2D observation sequence as input. The spatial embeddings are first encoded by a spatial Transformer and then fed to a dual temporal Transformer, named a multi-granularity temporal Transformer. Then holistic and local branches are designed to model complicated temporal dependencies elaborately. The outputs of dual branches are fused to obtain the final prediction. The $N_1\times$ represents the layer depth of HTT, and $N_2\times$ denotes the layer depth of JTTT.

ground truth joint, $\hat{Y}_k$ represents the estimated coordinate of the k_{th} joint. The loss function $\mathcal{L}$ aggregates all joint coordinate errors and calculates the average value.

3.4 Network Architecture

The above (Fig. 2) illustrates the framework of the 2D-3D pose regression network model, encompassing its structural composition and design pattern. Detailed descriptions of several modules are provided below.

Spatial Transformer Encoder. Given a series of 2D joint coordinates generated by CPN [2] generator or real 2D joint coordinates, the spatial transformer encoder (STE) utilizes a fully connected layer to elevate the dimension to 32. Subsequently, it incorporates learnable position encoding into the respective joint encoding, thereby creating a joint encoding feature denoted as C.

Holistic Temporal Transformer. The holistic temporal transformer (HTT), based on the output features of STE, utilizes one-dimensional convolution to convert high-dimensional features into low-dimensional ones S and employs positional information for time series modeling. To reduce redundancy between frames, a technique involving the addition of strides is utilized.

Joint Tracing Temporal Transformer. To capture the temporal dynamics between individual joints, a one-dimensional convolution is employed to increase the dimensionality of the low-dimensional features outputted by STE to 256 dimensions. The transformer architectures of both JTTT and HTT are similar, as they both utilize temporal information to learn and output features $G \in R^{J\times 1\times 256}$ of the central frame.

Regression Head. To regress the multi-granularity temporal features to 3D pose coordinates of the center frame. This framework adopts a parallel mode and concatenates the outputs of HTT and JTTT to produce a $H \in R^{1\times 256}$. Then a one-dimensional convolution operates on it and produces a $Y \in R^{J\times 3}$.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

The Human3.6M dataset is widely used for indoor 3D human pose estimation [32]. This dataset captures four camera views of actors performing various actions, including sitting, smoking, and walking. The dataset provides video sequences and corresponding 2D and 3D pose coordinates. Following previous works [3, 8], actors S1, S5, S6, S7, and S8 are included in the training set, while actors S9 and S11 are divided into the test sets. We evaluate the performance of our approach using two criteria: MPJPE and the aligned MPJPE (P-MPJPE). MPJPE calculates the average Euclidean distance (in millimeters) between predicted joints and real 3D joints, as detailed in protocol #1. P-MPJPE computes MPJPE using real 3D joint coordinates after applying rigid body transformation, such as translation, rotation, and scaling, to the output joint coordinates, as outlined in protocol #2. The MPI-INF-3DHP [33] dataset has both indoor and outdoor pose information, providing a broader perspective with action shots captured from 14 different angles. It consists of 8 actors, evenly split between male and female, each wearing two different outfits and performing 8 distinct movements, in which each movement lasts about one minute. The dataset contains real 3D coordinate information, and the experiments are divided according to [11], using MPJPE, percentage of correct key points (PCK) within 150mm, and area under the curve (AUC) evaluation metric.

4.2 Implementation Details

The model is implemented by the PyTorch framework and is trained on four NVIDIA GeForce RTX 3090 GPUs. We set the depth of the STE to 2, and each joint point dimension channel to 32, which corresponds to $N_1\times$=3 in JTTT and $N_2\times$=3 in HTT. During training, a batch size of 160 is used, and the initial learning rate is set to 0.01 with a decay rate per round of 0.95. A large decay is applied every five rounds, and the total number of epochs is set to 100. Adam with a weight decay of 1×10^{-6} is used. Additionally, we experimented with two activation functions, RELU and GLUE, and compared the results from both quantitative and qualitative perspectives. The lengths of input sequences following [10, 11] set to 27, 81, 243, and 351. To generate the 2D pose coordinates, we use a pre-trained CPN [37]. During training, the inputs consist of ground-truth 2D pose coordinates, while during testing, they generate 2D pose coordinates.

4.3 Baselines

We compare MOTT with twelve previous methods [2–11, 13–16, 23, 26, 34]. These methods include single-frame input and multi-frame input, in which [6, 13, 15, 16, 23] are based on monocular-pose estimation, and [2–5, 7–11, 26, 34] are based on video-pose estimation. For a

Table 1 Comparison experiments on Human3.6M based on protocol #1. The **red color and the blue color denote the best results and the second best results.**

Protocol #1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WaklT.	Avg.
Martinez et al. [13] ICCV'17	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang et al. [14] AAAI'18	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Lee et al. [15] ECCV'18 †	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Cai et al. [2] ICCV'19 †	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	**37.1**	37.1	39.4	48.8
Pavllo et al. [3] CVPR'19 †	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Lin et al. [4] BMVC'19 †	42.5	44.8	42.6	44.2	48.5	57.1	42.6	41.4	56.5	64.5	47.4	43.0	48.1	33.0	35.1	46.6
Xu et al. [6] CVPR'20 †	**37.4**	43.5	42.7	42.7	46.6	59.7	41.3	45.1	**52.7**	60.2	45.8	43.1	47.7	33.7	37.1	45.6
Liu et al. [7] CVPR'20 †	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Zeng et al. [16] ECCV'20 †	46.6	47.1	43.9	41.6	45.8	**49.6**	46.5	**40.0**	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
Wang et al. [5] ECCV'20 †	40.2	**42.5**	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
Chen et al. [34] TCSVT'21 †	41.4	43.5	40.1	42.9	46.6	51.9	41.8	42.3	53.9	60.2	45.4	**41.7**	46.0	31.5	32.7	44.1
Li et al. [10] TMM'22	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	**44.2**	43.0	44.2	30.0	30.2	43.7
Yu et al. [35] ICCV'23	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
MOTT †	39.0	43.1	**37.3**	**40.5**	**44.3**	51.8	**40.4**	40.4	56.3	**59.2**	**44.2**	42.0	**42.2**	**28.5**	**29.5**	**42.6**

Table 2 Comparison experiments on Human3.6M based on protocol #2.

Protocol #2	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WaklT.	Avg.
Martinez et al. [13] ICCV'17	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	45.0	38.0	43.1	47.7
Pavlakos et al. [21] CVPR'18	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Cai et al. [2] ICCV'19 †	35.5	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Lin et al. [4] BMVC'19 †	32.5	35.3	34.3	36.2	37.8	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo et al. [3] CVPR'19 †	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Xu et al. [6]CVPR'20 †	**31.0**	**34.8**	34.7	34.4	36.2	43.9	31.6	33.5	**42.3**	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu et al. [7]CVPR'20 †	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	**32.4**	37.0	25.2	27.2	35.6
Wang et al. [5] ECCV'20 †	32.9	35.2	35.6	34.4	36.4	42.7	**31.2**	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Wang et al. [34] TCSVT'21	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3	42.7	49.0	36.8	**32.4**	36.0	24.9	26.5	35.0
Zheng et al. [8] ICCV'21	32.5	**34.8**	32.6	34.6	35.3	**39.5**	32.1	32.0	42.8	48.5	**34.8**	**32.4**	35.3	24.5	26.0	34.6
Li et al. [10] TMM'2022	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
Yu et al. [35] ICCV'23	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
MOTT †	32.0	35.6	**30.9**	**33.7**	35.5	**41.2**	32.4	**31.8**	45.2	**48.3**	36.4	32.8	**34.4**	23.2	**24.4**	**34.5**

Table 3 Comparison experiments on Human3.6M based on protocol #1 with 2D ground-truth.

GT Protocol#1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WaklT.	Avg.
Pavllo et al. [3] CVPR'19	35.2	40.2	32.7	28.6	38.2	45.5	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Liu et al. [7] CVPR'20	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Zeng et al. [16] ECCV'20	34.8	32.1	28.5	30.7	31.4	36.9	35.6	30.5	38.9	40.5	32.5	31.0	29.9	22.5	24.5	32.0
Zheng et al. [8] ICCV'21	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Li et al. [9] CVPR'22	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
Ma et al. [26] ICMEW'22	29.4	30.1	27.5	27.4	30.5	32.8	32.3	29.5	33.4	37.0	29.6	29.1	29.2	23.9	24.5	29.8
San et al. [11] ECCV'22	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
Zhang et al. [36] CCDC'23	27.5	31.1	30.1	29.5	30.4	34.6	32.9	30.2	37.1	38.9	30.8	30.2	29.6	22.2	22.9	30.5
MOTT †	**25.6**	**27.3**	**25.6**	**25.4**	**27.8**	**30.3**	**28.5**	**24.9**	**35.8**	**37.2**	**27.5**	**27.8**	**24.8**	**18.5**	**19.7**	**27.1**

fair comparison, the inputs of the methods are 2d human body joint coordinates estimated by CPN [37] in Table 1 and 2. To address the uncertainty of input 2d coordinates, we used ground-truth 2D coordinates as input for comparison methods in Table 3.

4.4 Comparison Quantitative Results

Referring to previous methods [9–11], we verified the performance of MOTT on the two challenging datasets, Human3.6M and MPI-INF-3DHP.

Table 4 Model performance and model size on Human3.6M based on protocol #1.

Method	Params(M)	MPJPG (↓)
Zhang *et al.* [12] CVPR'22	33.7	42.4
Li *et al.* [9] CVPR'22	31.5	43.0
MOTT	8.9	42.6

Results on Human3.6M. Table 1 and Table 2 document the qualitative experimental results of different methods on the Human3.6M dataset when using the 2D coordinates generated by CPN. Table 3 records the experimental results when using real 2D coordinates as input. By comparing the data, it is observed that in Table 3, the MOTT method reduces the average error rate from 30.5 millimeters to 27.1 millimeters. For specific actions, there are significant increases in accuracy, with WalkDog increasing by 11.7%, Eating by 10.8%, Directions by 7%, and Discussions by 10%. Additionally, in both Table 1 and Table 2, the MOTT method outperforms baseline methods, indicating a significant improvement for simpler actions but less prominent improvements for more complex actions.

Furthermore, we compare the model parameter size and performance with some newer techniques with better results, as shown in Table 4. In cases where there is little difference in accuracy, our method has significantly fewer model parameters compared to others. MOTT achieves outstanding performance with a relatively lightweight model.

Results on MPI-INF-3DHP. We conducted experiments against the state-of-the-art on MPI-INF-3DHP dataset, as shown in Table 5. From the table, MOTT achieves a significant improvement over the method, San *et al.*, with the second-best results on MPJPE, outperforming by 10.3%. Since the sequence length in MPI-INF-3DHP is shorter than that in the

Human3.6M dataset, MOTT has a promising generalization capability in tackling datasets with different sequence lengths.

Table 5 Quantitative Comparison Results on MPI-INF-3DHP.

Method	PCK	AUC	MPJPE
Mehta *et al.* [38] 3DV'17 (F=1)	75.7	39.3	117.6
Pavllo *et al.* [3] CVPR'19 (F=81)	86.0	51.9	84.0
Lin and Lee. [4] BMVC'19 (F=25)	83.6	51.4	79.8
Zeng *et al.* [16] ECCV'20 (F=1)	77.6	43.8	-
Wang *et al.* [5] ECCV'20 (F=96)	86.9	62.1	68.1
Zheng *et al.* [8] ICCV'21 (F=9)	88.6	56.4	77.1
Chen *et al.* [34] TCSVT'21 (F=81)	87.9	54.0	78.8
San *et al.* [11] ECCV'22 (F=81)	**97.9**	75.8	32.2
Zhang *et al.* [36] CCDC'23 (F=81)	**97.9**	75.3	32.8
MOTT (F=81)	97.4	**77.2**	**28.9**

4.5 Ablation Study

The MOTT model consists of multiple components and modules. To verify the effectiveness of each component, we conducted the following experiments.

Effect of Multi-Granularity Temporal Transformer. The multi-granularity joint tracing transformer contains three components: the spatial transformer encoder (STE), the holistic temporal transformer (HTT), and the joint-tracing temporal transformer (JTTT). Using an input sequence of 81 frames, we study the influence of each component module on the overall structure. As shown in Table 6, MPJPE, Params, and Flops were used as evaluation indicators. To demonstrate the effectiveness of each module, we first compare an individual HTT module with a combined HTT and STE module. The results show that MPJPE decreases from 45.94 mm to 45.52 mm, with a 2.4% improvement in the prediction accuracy of the center frame. This indicates that the addition of STE is effective in learning the spatial topology information of the human body structure.

Table 6 Ablation studies on network components of MOTT.

STE	HTT	JTTT	Params(M)	FLOPs(G)	MPJPE (↓)
×	√	×	4.62	133.34	45.94
√	√	×	4.64	136.95	45.52
×	√	√	7.99	139.62	45.31
√	√	√	**8.00**	**143.23**	**44.67**

Similarly, to demonstrate the effectiveness of the JTTT module, we incorporate the JTTT module into the HTT base block, improving the model's MPJPE accuracy to 45.31. Finally, the three modules are integrated. From 6, it can be observed that the error rate of the combined model structure is reduced from 45.94 mm to 44.67 mm, with an accuracy improvement of 1.9%. Such controlled variable experiments effectively demonstrate the effectiveness of each module's integration in predicting temporal video frames.

Architecture Parameter Analysis. The selection and design of model parameters are vital to the performance of our model. As shown in Table 7, we discuss the impact of the number of layers of JTTT ($N_2\times$), depth of HTT ($N_1\times$), feature dimension of a single frame (embedding), the number of hidden layer channels (channels), the encoding feature dimension of JTTT (Dhid) on the model effect. In Table 7, bold red color represents the best result, and denote bold green color represents the worst result.

Table 7 Ablation study on different parameters of MOTT. We report the MPJPE.

Group	$N_2\times$	$N_1\times$	Embeddim	Channel	Dhid	MPJPE
1	1	2	32	256	512	45.58
	2	2	32	256	512	45.22
	3	2	32	256	512	44.67
	4	2	32	256	512	45.52
	5	2	32	256	512	44.78
	6	2	32	256	512	45.07
2	3	1	32	256	512	45.25
	3	**2**	**32**	**256**	**512**	**44.67**
	3	3	32	256	512	45.35
	3	4	32	256	512	45.29
	3	5	32	256	512	45.67
3	**3**	**2**	**32**	**256**	**512**	**44.67**
	3	2	32	512	1024	44.74
	3	2	32	512	2048	44.62
4	3	2	64	256	512	44.97
	3	2	64	512	1024	45.05
	3	2	64	512	2048	45.20
5	**3**	**2**	**128**	**256**	**512**	**45.59**
	3	2	128	512	1024	44.71
	3	2	128	512	2048	44.79
6	3	2	256	256	512	44.90
	3	2	256	512	1024	45.12
	3	2	256	512	2048	44.95

In the first set of parameter configurations, experimental results indicate that MOTT achieves optimal performance when $N_2\times$* is set to 3. Therefore, we fix $N_1\times$)* to 3 for the subsequent parameter analysis. To discuss the effectiveness of the depth, we analyze the performance of the MPJPE metric in the second set of experiments. We observe that the predicted 3D coordinates have the highest accuracy by setting Depth=2. From the third through sixth sets of experiments, we focus on testing the number of embedded channels and the dimension of joint points. MOTT performs best when the embedding size is set to 32 and the channel to 256. From the table, The optimum prediction result is 44.67 mm, and the error rate of the most unreasonable parameters in the first group is 45.59 mm, which improves the performance by 2.0%.

Impact of the Number of Frames. Unlike single-frame pose estimation, the number of input frames in video pose estimation is critical. To verify the effectiveness of the number of frames, we conducted experiments using varying numbers of input frames. Utilizing CPN [37] and ground-truth 2D sequences with different lengths, we evaluated the Mean Per Joint Position Error (MPJPE). The input 2D pose coordinates are obtained in two ways: one is generated by the CPN network as a pose generator, and the other is the actual 2D coordinates.

From Table 8, the accuracy rate of 81 frames is higher than that of 27 frames, and the error rate of 243 frames is also lower than that of 81 frames. MOTT achieves its best performance with 351 frames, demonstrating that MOTT has a solid ability to capture the correlation between long sequences of frames.

Table 8 Ablation study on inputs of MOTT.

Frames	27	81	243	351
w/ CPN detections	46.0	44.7	43.5	42.6
w/ gt 2D observations	36.2	33.62	28.0	27.1

4.6 Qualitative Results

Attention Visualization. We present attention visualization [10, 11] information of MOTT on subject S9 in the Human3.6M dataset. The visualization of multi-head self-attention intuitively displays the regions and parts

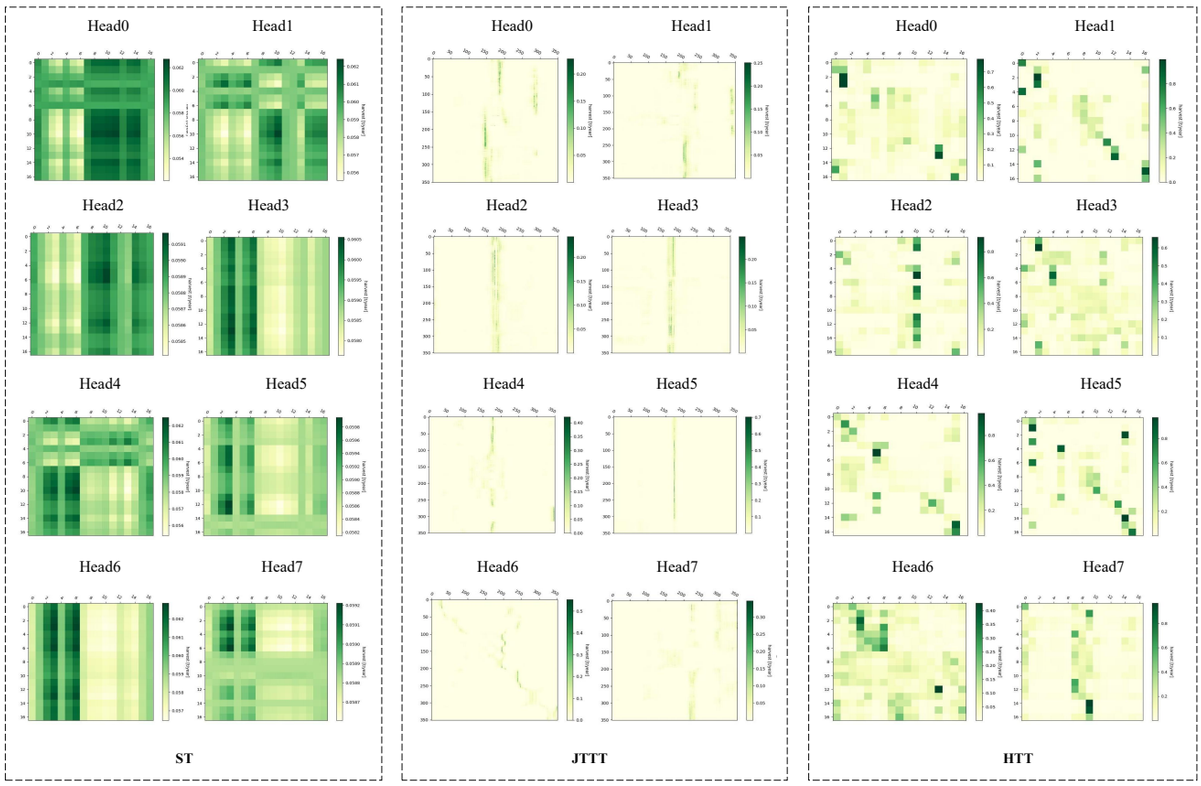

Fig. 3 Visualization of attention map of MOTT. We visualize the dependencies encoded by spatial transformer encoder (STE), holistic temporal transformer (HTT), and joint tracing temporal transformer (JTTT).

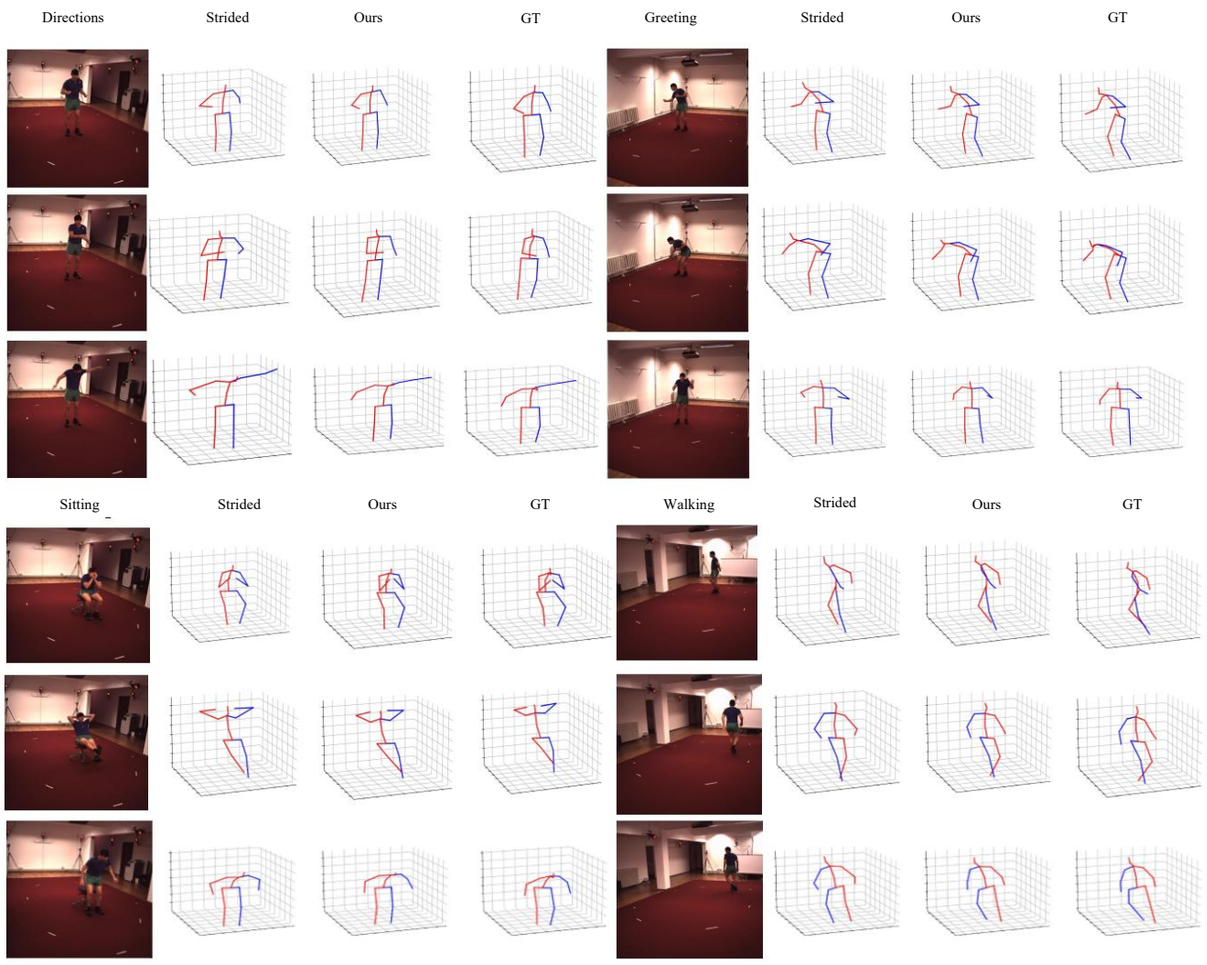

Fig. 4 Qualitative evaluation on Human3.6M. We compare the proposed MOTT against Strided Transformer [10] with four different actions. The input images, results of Strided Transformer, Our results, and the ground-truth 3D poses are shown side by side.

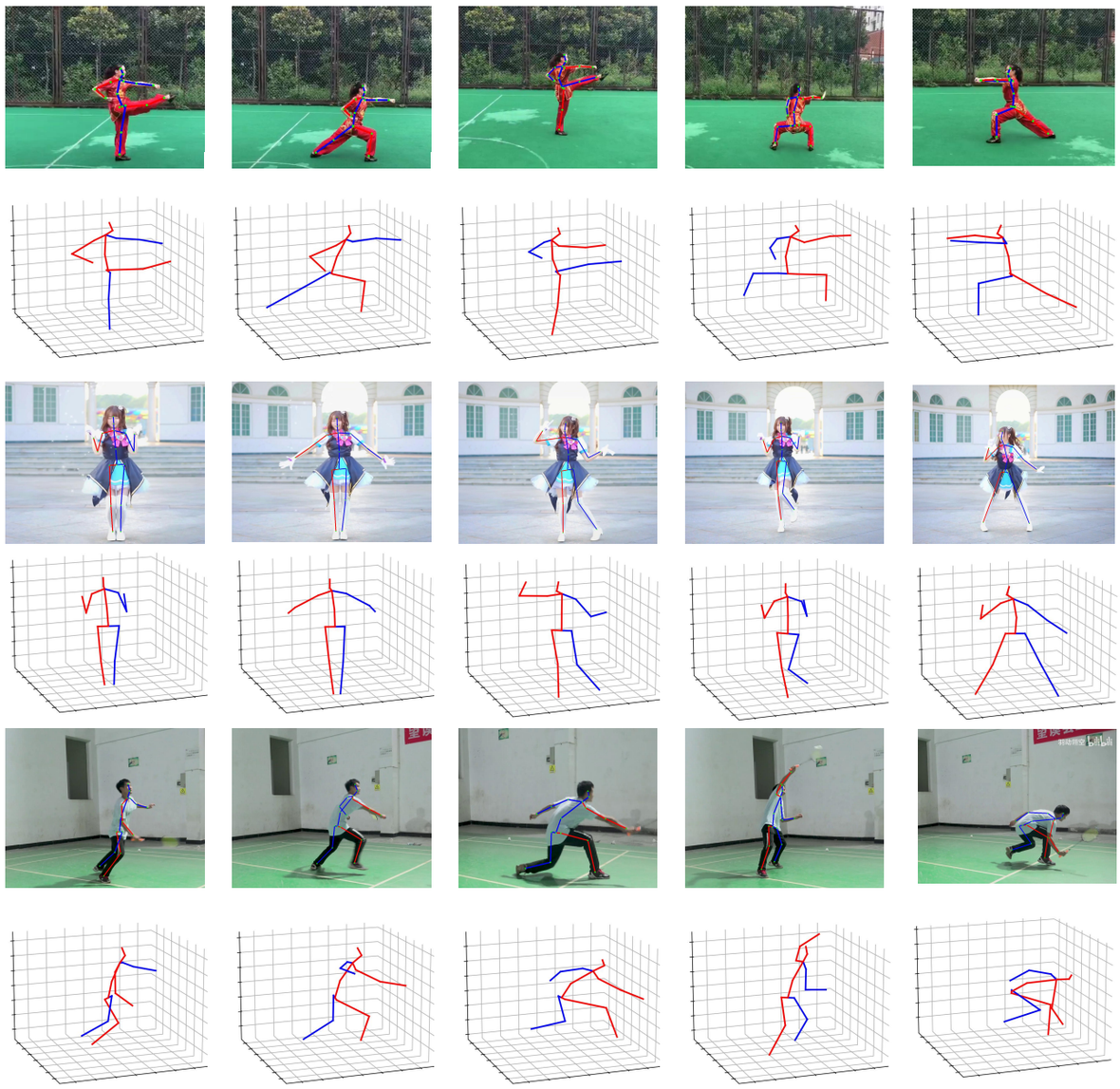

Fig. 5 Qualitative results on in-the-wild videos. We collect about 30 in-the-wild videos to evaluate the performance of our method in such an extreme case. It demonstrates that our MOTT performs well on this data.

that the model pays attention to. In Fig. 3, the 8 Heads in the STE part respectively represent the attention to the 17 joint points of the human body, with the x-axis representing the order of the joint points and the y-axis representing the weight of attention. We observe that heatmap 0 is mainly concentrated in 0 (Hip), 8 (thorax), 9 (neck), 10 (head), 11 (left-shoulder), 12 (left-elbow), 13 (left-wrist), 14 (right-shoulder), among others. Heatmap 2 focuses on several repeated concerns such as the 14 (right-shoulder), 15 (right-bow), and 16 (right-wrist). Heatmap 0 and heatmap 2 learn the key to the upper body of the human body's local correlation of points, similarly, heatmap 4 and heatmap 6 focus on joint points in the lower body such as the right-hip (1), right-knee (2), right-foot (3), left-hip (4), left-knee (5), and left-foot (6). The HTT module mainly focuses on the influence of surrounding frames on the center frame. The weight assignments after Transformer learning in different modules are visualized separately. The x-axis represents the input frame, and the y-axis represents the influence weight of adjacent frames. In the experiment, 351 frames are used as input frames, and the center of gravity of the heatmap is concentrated around the target frame. The JTTT module models the sequence of the same node learns the spatial structure features of the target output frame and pays attention to the pose features of the current node.

Dataset Visualization. To observe the performance of the model on different datasets, we conducted a qualitative comparison between MOTT and the current advanced methods. Fig. 4 shows that on the Human3.6M dataset, different methods are applied to the same dynamic prediction. Additionally, to verify the generalization of models to the wild dataset, we collected multiple daily action videos such as dancing, martial arts, badminton, etc. As shown in Fig. 5, the estimations of MOTT are closer to the truth than the baselines. For a fair comparison, we utilized CPN [37] as the 2D Pose generator.

5 Conclusion

In this work, we present a multi-granularity joint tracing Transformer (MOTT) approach based on 2D-3D lifting for 3D human pose estimation. It considers the body-joint correlation in video frames from multiple granularities, *i.e.*, the temporal correlation of holistic body joints and each single body joint. We design a multi-granularity temporal Transformer that models the holistic temporal relationship of all body joints and the local temporal relationship of every single joint from 2D image sequences. Extensive experimental results demonstrate the effectiveness of MOTT and indicate that the multi-granularity temporal Transformer is effective for human pose estimation.

References

[1] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392 (2020)

[2] Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2272–2281 (2019)

[3] Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2019)

[4] Lin, J., Lee, G.H.: Trajectory space factorization for deep video-based 3d human pose estimation. arXiv preprint arXiv:1908.08289 (2019)

[5] Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3d pose estimation from videos. In: European Conference on Computer Vision, pp. 764–780 (2020). Springer

[6] Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W.: Deep kinematics analysis for monocular 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 899–908 (2020)

[7] Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.-c., Asari, V.: Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5064–5073 (2020)

[8] 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)

[9] 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)

[10] Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia (2022)

[11] 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. arXiv preprint arXiv:2203.07628 (2022)

[12] Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: 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)

[13] Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2640–2649 (2017)

[14] Fang, H.-S., Xu, Y., Wang, W., Liu, X., Zhu, S.-C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[15] Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 119–135 (2018)

[16] Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: European Conference on Computer Vision, pp. 507–523 (2020). Springer

[17] Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3425–3435 (2019)

[18] Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7307–7316 (2018)

[19] Liu, M., Wang, W., Zhao, W.: Pva-gcn: point-voxel absorbing graph convolutional network for 3d human pose estimation from monocular video. Signal, Image and Video Processing, 1–15 (2024)

[20] Panda, A., Mukherjee, D.P.: Monocular 3d human pose estimation by multiple hypothesis prediction and joint angle supervision. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 3243–3247 (2021). IEEE

[21] Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7025–7034 (2017)

[22] Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)

[23] Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[24] Zhao, W., Tian, Y., Ye, Q., Jiao, J., Wang, W.: Graformer: Graph convolution transformer for 3d pose estimation. arXiv preprint arXiv:2109.08364 (2021)

[25] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)

[26] Ma, H., Lu, K., Xue, J., Niu, Z., Gao, P.: Local to global transformer for video based 3d human pose estimation. In: 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6 (2022). IEEE

[27] 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) (2018)

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[29] Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A.: Learning 3d human pose from structure and motion. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 668–683 (2018)

[30] Ma, H., Chen, L., Kong, D., Wang, Z., Liu, X., Tang, H., Yan, X., Xie, Y., Lin, S.-Y., Xie, X.: Transfusion: Cross-view fusion with transformer for 3d human pose estimation. arXiv preprint arXiv:2110.09554 (2021)

[31] Shuai, H., Wu, L., Liu, Q.: Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[32] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)

[33] 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 (2017). IEEE

[34] Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 198–209 (2021)

[35] Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.-h., 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)

[36] Zhang, K., Luan, X., Syed, T.H.S., Xiang, X.: Icrformer: An improving cos-reweighting transformer for 3d human pose estimation in video. In: 2023 35th Chinese Control and Decision Conference (CCDC), pp. 436–441 (2023). IEEE

[37] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)

[38] 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 camera. Acm transactions on graphics (tog) **36**(4), 1–14 (2017)