Trajectory-Aware Body Interaction Transformer for Multi-Person Pose Forecasting

Xiaogang Peng, Siyuan Mao, Zizhao Wu†
Department of Digital Media Technology, Hangzhou Dianzi University, Hangzhou, China
{pengxiaogang,siyuanmao,wuzizhao}@hdu.edu.cn

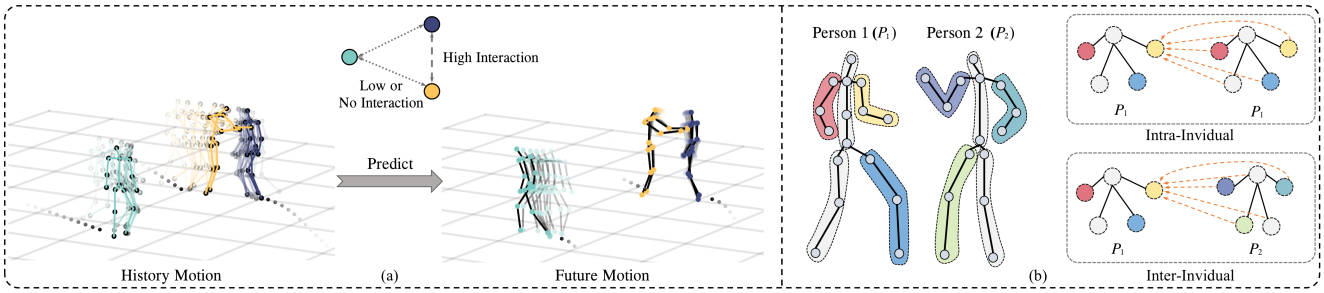

Figure 1. (a) In complex crowd scenarios, different people may interact with one another at varying levels (low and high interactions) and at different positions (*i.e.*, between near and far distances). (b) The illustration of our main idea on body part interactions. We divide the body joints into 5 parts, and the Intra-Individual branch is used to explore part relationships for each individual and the Inter-Individual branch aims to capture interaction dependencies of body parts between individuals. Our TBIFomer facilitates to model body part interactions for intra- and inter-individuals simultaneously.

Abstract

Multi-person pose forecasting remains a challenging problem, especially in modeling fine-grained human body interaction in complex crowd scenarios. Existing methods typically represent the whole pose sequence as a temporal series, yet overlook interactive influences among people based on skeletal body parts. In this paper, we propose a novel ***T****rajectory-Aware* ***B****ody* ***I****nteraction Trans****former*** *(****TBIFormer****) for multi-person pose forecasting via effectively modeling body part interactions. Specifically, we construct a Temporal Body Partition Module that transforms all the pose sequences into a Multi-Person Body-Part sequence to retain spatial and temporal information based on body semantics. Then, we devise a Social Body Interaction Self-Attention (SBI-MSA) module, utilizing the transformed sequence to learn body part dynamics for inter- and intra-individual interactions. Furthermore, different from prior Euclidean distance-based spatial encodings, we present a novel and efficient Trajectory-Aware Relative Position Encoding for SBI-MSA to offer discriminative spatial information and additional interactive clues. On both short- and long-term horizons, we empirically evaluate our framework on CMU-Mocap, MuPoTS-3D as well as synthesized datasets (6 $\sim$ 10 persons), and demonstrate that our method greatly outperforms the state-of-the-art methods.*

1. Introduction

Recent years have seen a proliferation of work on the topic of human motion prediction [4, 6, 7, 13, 24, 25, 28, 34], which aims to forecast future poses based on past observations. Similarly, understanding and forecasting human motion plays a critical role in the field of artificial intelligence and computer vision, especially for robot planning, autonomous driving, and video surveillance [8, 14, 21, 44]. Although encouraging progress has been achieved, the current methods are mostly based on local pose dynamics forecasting without considering global position changes of body joints (global body trajectory) and often tackle the problem of single humans in isolation while overlooking human-human interaction. Actually, in real-world scenarios, each person may interact with one or more people, ranging from low to high levels of interactivity with instantaneous and deferred mutual influences [2, 31]. As illustrated in Fig. 1 (a), two individuals are pushing and shoving with high interaction, whilst a third individual is strolling with no or low interaction. Thus, accurately forecasting pose dynamics and trajectory and comprehensively considering complex social interactive factors are imperative for understanding human behavior in multi-person motion prediction. However, existing solutions do not efficiently address these challenging factors. For example, Guo *et al*. [15] propose a collaborative prediction task and perform future motion prediction for only two interacted dancers, which inevitably ignores low interaction influence on one's future behavior. Wang *et al*. [39] use local and global Transformers to learn indi-

†Corresponding author.

vidual motion and social interactions separately in a crowd scene. The aforementioned methods ignore the interactive influences of body parts and only learn temporal and social relationships without modeling fine-grained body interaction, which makes it difficult to capture complex interaction dependencies.

To solve this issue, we propose a novel Transformer-based framework, termed TBIFormer, which consists of multiple stacked TBIFormer blocks and a Transformer decoder. In particular, each TBIFormer block contains a Social Body Interaction Multi-Head Self-Attention (SBI-MSA) module, which aims at learning body part dynamics across inter- and intra-individuals and capturing fine-grained skeletal body interaction dependencies in complex crowd scenarios as shown in Fig. 1 (b). More specifically, SBI-MSA learns body parts dynamics across temporal and social dimensions by measuring motion similarity of body parts rather than pose similarity of the entire body. In addition, a Trajectory-Aware Relative Position Encoding is introduced for SBI-MSA as a contextual bias to provide additional interactive clues and discriminative spatial information, which is more robust and accurate than the Euclidean distance-based spatial encodings.

In order to feed the TBIFormer a pose sequence containing both temporal and spatial information, an intuitive way is to retain body joints in time series. However, this strategy will suffer from noisy joints caused by noisy sensor inputs or inaccurate estimations. In this work, we propose a Temporal Body Partition Module (TBPM) that, based on human body semantics, transforms the original pose sequence into a new one, enhancing the network's capacity for modeling interactive body parts. Then, we concatenate the transformed sequences for all people one by one to generate a Multi-Person Body Part (MPBP) sequence for input of TBIFormer blocks, which enables the model to capture dependencies of interacting body parts between individuals. TBIFormer makes MPBP sequence suitable for motion prediction by utilizing positional and learnable encodings to indicate to whom each body part and timestamp belongs.

Finally, a Transformer decoder is used to further consider the relations between the current and historical context across individuals' body parts toward predicting smooth and accurate multi-person poses and trajectories. For multi-person motion prediction (with $2 \sim 3$ persons), we evaluate our method on multiple datasets, including CMU-Mocap [9] with UMPM [35] augmented and MuPoTS-3D [30]. Besides, we extend our experiment by mixing the above datasets with the 3DPW [38] dataset to perform prediction in a more complex scene (with $6 \sim 10$ persons). Our method outperforms the state-of-the-art approaches for both short- and long-term predictions by a large margin, with $14.4\% \sim 16.5\%$ accuracy improvement for the short-term (≤ 1.0s) and $6.5\% \sim 18.2\%$ accuracy improvement for the long-term ($1.0\text{s} \sim 3.0\text{s}$).

To summarize, our key contributions are as follows: 1) We propose a novel Transformer-based framework for effective multi-person pose forecasting and devise a Temporal Body Partition Module that transforms the original pose sequence into a Multi-Person Body-Part sequence to retain both temporal and spatial information. 2) We present a novel Social Body Interaction Multi-Head Self-Attention (SBI-MSA) that learns body part dynamics across inter- and intra-individuals and captures complex interaction dependencies. 3) A novel Trajectory-Aware Relative Position Encoding is introduced for SBI-MSA to provide discriminative spatial information and additional interactive clues. 4) On multiple multi-person motion datasets, the proposed TBIFormer significantly outperforms the state-of-the-art methods.

2. Related Work

2.1. Single-Person Pose Forecasting

Predicting human motion offers enormous promise for surveillance, autonomous driving, and human-robot interaction. Although recurrent neural networks (RNNs) have shown advantages in processing this typical sequence-to-sequence problem [13, 18, 29], discontinuity and error accumulation often happen due to the frame-by-frame prediction manner. To address these issues, some feed-forward networks such as graph convolution networks (GCNs) and temporal convolution networks (TCNs) are used to explore spatial and temporal dependencies [7, 10, 12, 23, 25]. Besides, Mao *et al*. [26] introduce an attention-based feed-forward network to capture the similarity between the current motion context and the historical motion sub-sequences and process the result via GCNs for long-term prediction. All the above methods only model local pose dynamics, ignoring global body translation and inter-individual body interaction. However, learning both local and global pose dynamics and modeling fine-grained human-human interaction are essential for comprehending human behavior in a complex 3D environment [2, 39].

2.2. Multi-Person Pose Forecasting

In order to address fine-grained human-human interaction, some recent approaches are proposed for multi-person pose and trajectory forecasting. For example, Adeli *et al*. [1] propose to combine scene context and use graph attention networks to model interaction between humans and objects. Guo *et al*. [15] present a collaborative prediction task and use a two-branch attention network for the prediction of two interacted persons. Wang *et al*. [39] present a Transformer-based framework to forecast multi-person motion in a scenario with more people. Furthermore, this method produces unrealistic poses since they solely concen-

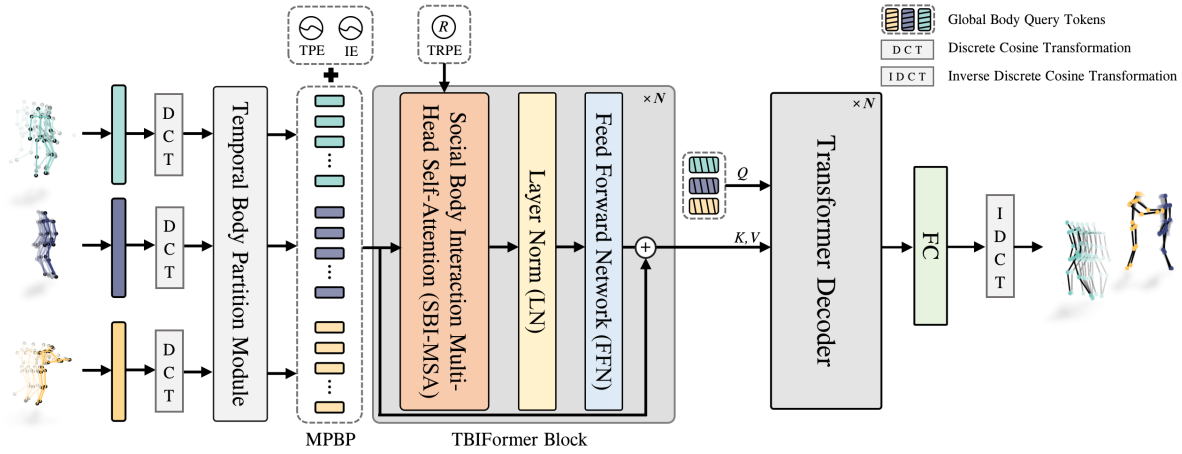

Figure 2. Overview of the proposed TBIFormer framework. Given the observed pose sequences of 3 persons, TBIFormer transforms them into displacement sequences as input and then forecasts future poses for each person. At the head and tail of TBIFormer, we adopt a Discrete Cosine Transformation (DCT) [3] that discards the high-frequency information for a more compact representation in displacement trajectory space [27].

trate on individual and social modeling in the time dimension. Despite the novelty of these works, skeletal body interaction between individuals is not captured effectively. In this work, we propose TBIFormer that learns skeletal body part dynamics for intra- and inter-individuals to effectively capture complex interaction dependencies.

2.3. Multi-Person Social Interaction

Pedestrian trajectory prediction is a representative issue for multi-person social interaction. Existing methods for the task can be categorized based on how they model time and social dimensions. RNNs [16] and Transformers [37] are the preferred models [5, 17, 43] to process the trajectory sequence for temporal modeling, and graph neural networks (GNNs) [20] are often adopted as social models for interaction modeling [19, 22, 41, 42]. While performing well, these studies only focus on individuals' global movement without modeling detailed human joint dynamics. In this work, we investigate our TBIFormer to consider fine-grained human-human interaction via modeling skeletal body part dynamics among individuals and predict future motion for $3 \sim 10$ persons in 3D scenes.

3. Method

In this section, we introduce our Trajectory-Aware Body Interaction Transformer (TBIFormer), which contains multiple stacked TBIFormer blocks and a Transformer decoder followed by fully connected layers, as shown in Fig. 2. Each TBIFormer block has a Social Body Interaction Multi-Head Self-Attention (SBI-MSA) module for modeling body part interactions across temporal and social dimensions. The proposed TBIFormer is also equipped with a Temporal Body Partition Module (TBPM), which aims to better learn body parts' spatial and temporal information within the skeletal sequences. In addition, temporal positional encoding, person identity encoding, and trajectory-aware relative position encoding are introduced to preserve time, identity, and discriminative spatial information. In the following, the problem definition and our key modules are described in detail.

3.1. Problem Definition

Supposing the observed skeletal poses from person p are $X^p_{1:T+1} = \{x^p_1, x^p_2, ..., x^p_{T+1}\}$ with $T+1$ frames, where $p = 1, 2, ...P$. For simplicity, we omit subscript p when p only represents an arbitrary person, *e.g.*, taking $x^p_{1:t}$ as $x_{1:t}$. Instead of absolute joint positions in the world coordinate, we use $y_i = x_{i+1} - x_i$ to obtain instantaneous pose displacement at time i, which will provides more valuable dynamics information [34, 39]. The whole displacement sequence is defined as $Y_{1:T} = \{y_1, y_2, ..., y_T\}$. Given the displacement sequence $Y_{1:T}$ of each person, our goal is to predict the N frames of future displacement trajectory $Y_{T+1:T+N}$ and transform it back to the pose space $X_{T+2:T+N+1}$.

3.2. Temporal Body Partition Module

To better retain both spatial and temporal information of the skeleton sequences, we propose TBPM that first transforms the pose sequence of each person into a sequence that contains body parts during a short time period. Then, TBPM concatenates all the transformed individuals' sequences into a Multi-Person Body-Part (MPBP) sequence for the following Transformers. There are three primary processes in TBPM, *i.e.*, partition, projection and concate-

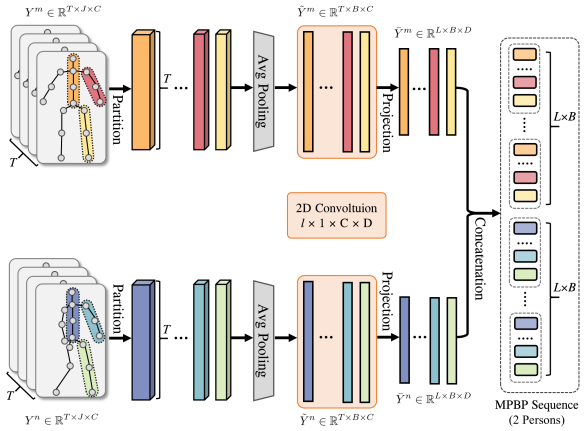

Figure 3. The illustration of the proposed Temporal Body Partition Module (TPBM). TPBM performs three main operations (*i.e.*, partition, projection, and concatenation) on the input pose sequence to generate a Multi-Person Body-Part (MPBP) sequence.

nation, which are described below.

Partition. Given the displacement sequence of the p-th person $Y^p \in \mathbb{R}^{T\times J\times C}$, where T and J represent the numbers of frames and joints, respectively, and $C = 3$ represents the dimension of the 3D coordinates, we first divide the sequence into $B = 5$ body parts (*e.g.* left and right arms, left and right legs, core torso) based on natural human skeletal structure, and then down-sample each body part by average-pooling. After the above operations, the sequence is represented as $\tilde{Y}^p \in \mathbb{R}^{T\times B\times C}$.

Projection. The goal of the projection operation is to initially extract spatial and temporal information. Specifically, we use 2D convolution with a kernel size of $l \times 1$ on $\tilde{Y}^p$ to obtain 2D feature map $\overline{Y}^p \in \mathbb{R}^{L\times B\times D}$, where $L = \lfloor (T - l + 1)/stride \rfloor$ and D denote the number of output channels. We denote "$padding$" and "$stride$" as the padding size and stride size of the convolutional filter.

Concatenation. Following projection, the encoding of all the B body parts are concatenated for all the L timesteps to form a new sequence with the length of $U = L \times B$. Next, we concatenate the sequences of all the P persons one by one for a merged Multi-Person Body-Part (MPBP $\in \mathbb{R}^{M\times D}$) sequence, where M denotes $P \times U$. MPBP sequence allows our TBIFormer to learn individuals' body part dynamics across temporal and social dimensions.

3.3. Temporal Positional and Person Identity Encoding

Similar to the original Transformer [37], we apply sinusoidal positional encoding to convey to TBIFormer the timestep associated with each element in the MPBP sequence. Instead of encoding the position of each element based on index in the whole MPBP sequence, we first compute timestamp features based on the timesteps of each person and obtain temporal positional encoding $\tau_p \in \mathbb{R}^{T\times d_\tau}$, where d_τ is the feature dimension of the timestamp. Then we utilize the interleaved repeating function to repeat the encoding elements for B body parts and concatenate the encoding of all individuals. The final temporal positional encoding (TPE) is formulated as $\hat{\tau} \in \mathbb{R}^{M\times d_\tau}$.

To provide identity information of each individual in the MPBP sequence, we also inject a learnable person identity encoding $\nu \in \mathbb{R}^{M\times d_\nu}$, indicating which individual each element belongs to, where d_ν denotes the feature dimension. Notably, the identity encoding (IE) is randomly initialized and repeated for the time and body parts using the same repeating method for TPE.

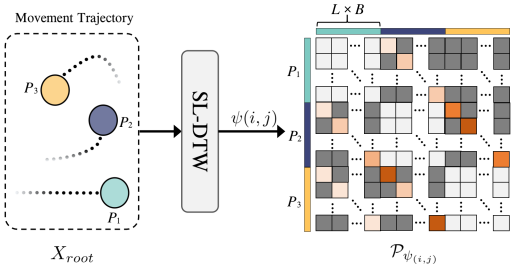

Figure 4. The overview of the proposed Trajectory-Aware Relative Position Encoding (TRPE).

3.4. Trajectory-Aware Relative Position Encoding

An instinctive assumption is that the closer a few people are, the higher the interaction they may have. Yet, in the complex crowd situation, one person may have their back turned to a nearby individual with no interaction, or, as depicted in Fig. 1, a person may just pass by two interacting individuals, exhibiting low interaction yet close proximity. Therefore, Euclidean distance-based spatial position encodings struggle to provide discriminative spatial information and distinguish individuals who are actually interacting.

In this paper, our observation is that people interacting in 3D space tend to move in the same or face-to-face direction, as opposed to the deviated direction. The main challenge is that directly calculating body orientation of the human skeleton data and the angle between the individuals is tedious and costly. To solve it, we find that movement trajectories can also provide vital information and circumvent the aforementioned limitations. Therefore, we propose a novel Trajectory-Aware Relative Position Encoding (TRPE) by measuring the similarity of movement trajectories, which can aggregate both corresponding movement pattern and spatial information. Dynamic Time Warping (DTW) [32, 33] is a more robust method to measure trajectory (series) similarity than Euclidean distance. In this work, we employ an efficient and differentiable algorithm

variant called Soft-DTW [11], which can be defined as,

$$D(i,j) = min\{D(i,j-1), D(i-1,j), D(i-1,j-1)\} + \delta(i,j), \quad (1)$$

where $D(i,j)$ denotes the shortest distance between subsequence $S1 = (s_1, s_2, ..., s_i)$ and $S2 = (s_1, s_2, ..., s_j)$ and $\delta(\cdot,\cdot)$ is differentiable cost function. In order to dynamically obtain trajectory similarity according to a certain timestep as opposed to complete timestamps, we propose a Shifted Local DTW (SL-DTW) mechanism based on Soft-DTW. Similar to the convolution operation, SL-DTW calculates the similarity between individuals at a specific window size and shifts step-by-step, which will provide more precise relative information. See Algorithm 1 for a detailed description of the SL-DTW process.

Algorithm 1 Shifted Local DTW mechanism (SL-DTW)

Input: The root trajectory sequence of person m and person n, $X_r^m = (x_{r,1}^m, x_{r,2}^m, ..., x_{r,T}^m)$ and $X_r^n = (x_{r,1}^n, x_{r,2}^n, ..., x_{r,T}^n)$; The size of local window and shift stride, l and $stride$; The length of input sequence, T;

Output: The trajectory similarity $D^{<m,n>}$ between person m and n;

$D^{<m,n>} = [\,]$

for $i = 0; i < \lfloor (T-l+1)/stride \rfloor; i += stride$ **do**

  $D^{<m,n>} = stack(D^{<m,n>}, D(x_{(r,i+l)}^m, x_{(r,i+l)}^n))$

end for

Given the trajectory similarity distance $\tilde{D} \in \mathbb{R}^{P\times L}$ among P persons, we need to map the distance to an integer set for relative position encoding. The common way to address this issue is the clip function: $h(\tilde{D}) = max(-\beta, min(\beta, \tilde{D}))$, which inevitably eliminates the context of long-distance relative position. Hence, we alternatively use the piecewise function [40] $g(\cdot)$ that maintains long-range information for indexing relative distances to corresponding encodings, and then define the indexed matrix through the SL-DTW distance as follows:

$$\psi(i,j) = \begin{cases} g(\eta), & \iota \neq \kappa, m \neq n, \\ g(0), & m = n, \\ g(\tilde{D}_{(\iota,\kappa)}^{<m,n>}), & \iota = \kappa, m \neq n, \end{cases} \quad (2)$$

where ι and κ denote different timesteps from different person. In Eq. (2), to reduce additional computation, we ignore relations between person m and n on the condition of $\iota \neq \kappa$ and input a larger value η in $g(\cdot)$ instead. The piecewise index function is presented as

$$g(e) = \begin{cases} [e], & |e| \leq \alpha, \\ sign(e) \times & |e| > \alpha, \\ min(\beta, [\alpha + \frac{\ln(|e|/\alpha)}{\ln(\gamma/\alpha)}(\beta-\alpha)]), & \end{cases} \quad (3)$$

where $[\cdot]$ is a round operation, $sign()$ determines the sign of a number, *i.e.*, returning 1 for positive input, -1 for negative, and 0 for otherwise. α controls the piecewise point, β limits the output in the range of $[-\beta, \beta]$, and γ tunes the curvature of the logarithmic part.

Finally, as shown in Fig. 4, we embed the indexed matrix $\psi(i,j)$ of trajectory similarity as our TRPE $\mathcal{P}_{\psi(i,j)} \in \mathbb{R}^{M\times d_z}$ and denote $M = P \times L \times B$, which are shared throughout all attention layers of SBI-MSA.

3.5. SBI-MSA Module

In each TBIFormer block, we aim to construct a Social Body Interaction Multi-Head Self-Attention (SBI-MSA) module to effectively model body part dynamics for inter- and intra-individual. Given the motion features extracted by TBPM, SBI-MSA, based on motion-wise attention computation, can further optimize pose dynamics and capture complex body interaction dependencies among individuals. Let $H = [h_1, ..., h_n] \in \mathbb{R}^{n\times d}$ denotes the input representation for attention module, where d is the hidden dimension. SBI-MSA takes as input keys K, queries Q and values V, each of which is projected by the corresponding parameter matrix $W_Q \in \mathbb{R}^{d\times d_z}$, $W_K \in \mathbb{R}^{d\times d_z}$ and $W_V \in R^{d\times d_z}$. The output of SBI-MSA is computed as

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V, \quad (4)$$

$$\text{SBI-MSA}(Q,K,V) = softmax(A)V. \quad (5)$$

We integrate the TRPE $\mathcal{P}_{\psi(i,j)}$ on the attention map to consider the interaction between individual dynamics features and spatial clues across temporal and social dimensions. Denoting A_{ij} as the (i,j)-element of the Query-Key product matrix A, we have

$$A_{ij} = \frac{Q_i \cdot K_j + b_{i,j}^{\text{TRPE}}}{\sqrt{d_z}}, \quad (6)$$

$$b_{i,j}^{\text{TRPE}} = Q_i \cdot \mathcal{P}_{\psi_{(i,j)}}, \quad (7)$$

where $b_{i,j}^{\text{TRPE}}$ is a contextual bias for the attention map.

3.6. Transformer Decoder

As illustrated in Fig. 2, we concatenate joint coordinates of the last observed sub-sequence (length = l) from each person for all the body joints and down-sample them on time dimension by 1D Convolution (kernel size = l) as global body query tokens. Key and value tokens are the output of the TBIFormer block. We utilize a standard Transformer decoder [37] to encode the relations between the current (queries) and historical context (keys) across individuals. At the end of the decoder, we adopt two fully connected (FC) layers followed by an Inverse Discrete Cosine Transformation (IDCT) [3] to generate the future motion trajectory $X_{T+2:T+N+1}$ for each individual.

		CMU-Mocap (UMPM) (3 persons)				MuPoTS-3D (2 ∼3 persons)				Mix1 (6 persons)				Mix2 (10 persons)			
	Method	0.2s	0.6s	1.0s	Overall	0.2s	0.6s	1.0s	Overall	0.2s	0.6s	1.0s	Overall	0.2s	0.6s	1.0s	Overall
JPE	HRI [26]	49	130	207	129	81	211	323	205	51	141	233	142	52	140	224	139
	MSR [12]	53	146	231	143	79	222	374	225	49	132	220	134	60	153	243	152
	MRT* [39]	36	115	192	114	78	225	349	217	37	122	212	124	38	126	214	126
	Ours*	**30**	**109**	**182**	**107**	**66**	**200**	**319**	**195**	**34**	**121**	**209**	**121**	**34**	**118**	**198**	**117**
APE	HRI [26]	41	97	130	89	70	136	174	127	38	92	122	84	41	100	133	91
	MSR [12]	46	106	137	96	71	148	190	136	41	92	120	84	48	110	148	102
	MRT* [39]	36	108	159	101	71	166	217	151	36	109	166	104	38	115	178	110
	Ours*	**27**	**84**	**118**	**76**	**60**	**132**	**170**	**121**	**28**	**81**	**113**	**74**	**30**	**89**	**124**	**81**
FDE	HRI [26]	31	90	158	93	63	173	279	172	37	107	192	112	35	101	177	104
	MSR [12]	29	94	175	99	58	184	335	192	29	91	169	96	38	113	185	112
	MRT* [39]	27	88	157	91	59	187	309	185	29	100	189	106	29	98	185	104
	Ours*	**18**	**72**	**133**	**74**	**49**	**163**	**277**	**163**	**23**	**89**	**168**	**93**	**21**	**81**	**151**	**84**

Table 1. Results of JPE, APE and FDE (in mm) on different datasets. We compare our method with the previous SOTA methods for short-term and long-term predictions. Best results are shown in boldface. (* means multi-person motion prediction method.)

	JPE			**APE**			**FDE**		
Method	1.0s	2.0s	3.0s	1.0s	2.0s	3.0s	1.0s	2.0s	3.0s
HRI [26]	134	229	349	99	133	161	93	177	295
MSR [12]	134	256	371	97	142	165	92	204	316
MRT* [39]	148	256	352	130	187	218	109	216	315
Ours*	**118**	**225**	**329**	**89**	**132**	**152**	**78**	**172**	**273**

Table 2. Results of JPE, APE and FDE (in mm) on CMU-Mocap (UMPM) dataset. We compare our method with the previous SOTA methods for long-term prediction (1.0s $\sim$ 3.0s). Best results are shown in boldface. (* means multi-person motion prediction method.)

3.7. Loss Function

We use a reconstruction loss based on the Mean Per Joint Position Error (MPJPE) for optimization. In particular, for one training sample, the loss is represented as

$$L_{rec} = \frac{1}{J * N} \sum_{i=N+1}^{T+N} \sum_{j=1}^{J} ||\hat{y}_{i,j} - y_{i,j}||^2, \tag{8}$$

where $\hat{y}_{j,t}$ and $y_{j,t}$ are ground-truth and estimated pose displacement at time i. J represents the number of body joints.

4. Experiments

4.1. Implementation Details

We implement our framework in PyTorch, and the experiments are performed on Nvidia GeForce RTX 3090 GPU. We train our model for 50 epochs using the ADAM optimizer with a batch size of 32, a learning rate of 0.0003, and a dropout of 0.2. For the TBPM, the kernel size and stride of 2D convolutional filter are 10×1 and $stride = 1$, and $padding = 0$. The parameters in TRPE are: $\alpha = 1, \beta = 9, \gamma = \eta = 2000$. The dimensions d_z of keys, queries, and values in TBIFormer block and Transformer decoder are all set to 64, and the hidden dimension d of feed-forward layers is 1024. There are 3 stacked TBIFormer blocks and attention layers with 8 heads in the TBIFormer and Transformer decoder.

4.2. Datasets

To verify the effectiveness of TBIFormer, we run experiments on the CMU-Mocap (UMPM) dataset, which merges UMPM [35] into CMU-Mocap [9] for dataset expansion. Mix1 and Mix2 are blended by CMU-Mocap, UMPM, 3DPW [38], and MuPoTs-3D [30] datasets. We evaluate all the methods for generalization ability by testing on the MuPoTS-3D (2 $\sim$ 3 persons), Mix1 (6 persons), and Mix2 (10 persons) datasets with the model only trained on the CMU-Mocap (UMPM) dataset. Please refer to the appendix for a thorough explanation of why we do dataset expansion and the processing detail of mixing datasets.

4.3. Metrics of Evaluation

JPE Metric. We use Joint Position Error (JPE) based on Mean Per Joint Position Error (MPJPE) to measure the poses of all the individuals, including body trajectory:

$$\text{JPE}(X, \hat{X}) = \frac{1}{P \times J} \sum_{i=1}^{P} \sum_{j=1}^{J} ||X_j^i - \hat{X}_j^i||^2, \tag{9}$$

where P and J are the numbers of people and joints. X_j^i and $\hat{X}_j^i$ are the estimated and ground-truth positions of the joint j for person i.

APE Metric. We remove global movement and use Aligned mean per joint Position Error (APE) to measure pure pose position error:

$$\text{AME}(X, \hat{X}) = \text{JPE}(X - X_r, \hat{X} - \hat{X}_r), \tag{10}$$

where X_r and $\hat{X}_r$ are the estimated and ground-truth root positions of human body.

	JPE			APE			FDE		
Method	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s
w/o TBPM	32	117	195	28	87	123	21	76	142
w/o IE, TRPE	31	113	188	**27**	85	120	19	74	138
w/o TRPE	31	112	186	**27**	85	119	19	73	136
TRPE $\rightarrow$ EuPE	40	118	191	34	89	121	20	80	139
w/o SBI-MSA	40	128	208	29	92	129	27	85	151
Full	**30**	**109**	**182**	**27**	**84**	**118**	**18**	**72**	**133**

Table 3. Ablation studies on different components of TBIFormer. Our full method and its variants are evaluated on the CMU-Mocap (UMPM) in JPE metric.

	JPE			APE		
Method	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s
HRI [26]	51 $\uparrow$ 2	134 $\uparrow$ 4	212 $\uparrow$ 5	41 $\uparrow$ 0	98 $\uparrow$ 1	132 $\uparrow$ 2
MSR [12]	55 $\uparrow$ 2	149 $\uparrow$ 3	238 $\uparrow$ 7	46 $\uparrow$ 0	106 $\uparrow$ 0	136 $\downarrow$ 1
MRT* [39]	38 $\uparrow$ 2	124 $\uparrow$ 9	203 $\uparrow$ 9	49 $\uparrow$ 13	142 $\uparrow$ 34	223 $\uparrow$ 64
Ours*	31 $\uparrow$ 1	111 $\uparrow$ 2	184 $\uparrow$ 2	28 $\uparrow$ 1	85 $\uparrow$ 1	120 $\uparrow$ 2

Table 4. Results on effects of random person permutation in input. All the methods are evaluated on the CMU-Mocap (UMPM) in JPE and APE metrics. The red values represent floating errors.

FDE Metric. We also adopt the root position to evaluate the global movement of each person using a typical trajectory prediction metric: Final Displacement Error (FDE). The formula is described as follows:

$$\mathrm{FDE}(X_r, \hat{X}_r) = ||X_{r,N} - \hat{X}_{r,N}||^2, \tag{11}$$

where $X_{r,N}$ and $\hat{X}_{r,N}$ are the estimated and ground-truth root position of final pose at N-th predicted timestamp.

4.4. Baselines

We choose 3 code-released state-of-the-art (SOTA) approaches as baselines, including two single-person based methods: HRI [26] and MSR [12], and a recently released multi-person based method called MRT [39]. HRI [26] is an attention-based network, and MSR [12] is a GCN-based method, which both allow absolute coordinates as input. For short-term prediction, we train all these models with 50 frames (2.0s) of input and 25 frames (1.0s) of forecasting and evaluate on the 4 datasets. For long-term prediction, using the protocols in MRT [39], we set 15 frames (1.0s) of history as input to predict the future 45 frames (3.0s).

4.5. Results

To validate the prediction performance of TBIFormer, we follow the setting of the most single-person methods [12, 26] to show the quantitative and qualitative results of short- and long-term predictions, and compare our method with the baselines.

Quantitative Results. Table 1 reports the results of JPE, APE and FDE on the 4 different datasets. Our TBIFormer significantly outperforms the baselines in prediction accuracy. We achieve up to 13% $\sim$ 27% improvement when compared to the previous single-person-based methods and achieve up to 13% $\sim$ 16% improvement compared to the multi-person-based method. It can be noticed that MRT [39] performs poorly in the APE metric due to the lack of spatial modeling of the human skeleton. Besides, we report the results of long-term prediction (1.0s $\sim$ 3.0s) in Tab. 2. Our method consistently outperforms the baselines in the 3 metrics.

Qualitative Results. Figure 5 shows some examples of our visualization results compared to the baselines and the ground truth. The results of HRI [26] and MSR [12] show that they tend to converge to a static pose in the long-term predictions. Due to the deficiency in spatial modeling of the human body, MRT [39] generates some distort poses. By contrast, our method generates more plausible 3D human motion in practice, which is much closer to the ground truth than others. More visualization results are supplemented in the appendix.

4.6. Ablation Studies

We further conduct extensive ablation studies on CMU-Mocap (UMPM) to investigate the contribution of key technical components in TBIFormer, with results in Tab. 3. For more ablation results about the model, please refer to the appendix.

Effectiveness of TBPM. The TBPM constructs a sequence containing both temporal and spatial information for human poses. When it is removed and joint coordinates are directly concatenated for body joints in a pose sequence, TBIFormer cannot learn body part dynamics, and we can observe a significant performance decrease.

Effectiveness of IE and TRPE. Person identity encoding (IE) allows our method to distinguish element types in the MPBP sequence (*i.e.*, inform each token about identity information). After eliminating IE, the model's overall performance has decreased marginally. Trajectory-aware relative position encoding (TRPE) provides ample spatial and interactive clues for the model. When we remove TRPE, the performance drops substantially. In addition, as shown from (TRPE $\rightarrow$ EuPE) in Tab. 3, even after replacing TRPE with Euclidean distance-based position encoding, the performance is still sub-optimal. In FWe also provide t-SNE visualization [36] to demonstrate discriminative power between TRPE and SE (Euclidean-based encoding) in MRT [39]. Apparently, our model equipped with TRPE can obtain more accurate and compact representations.

Effectiveness of SBI-MSA. The goal of the SBI-MSA is to learn body part dynamics across temporal and social dimensions. As illustrated in the final row of Tab. 3, if the SBI-MSA is substituted with a standard self-attention module, our model only learns motion features for each person separately, resulting in poorer long-term performance.

Effects of Random Person Permutation. To ensure that

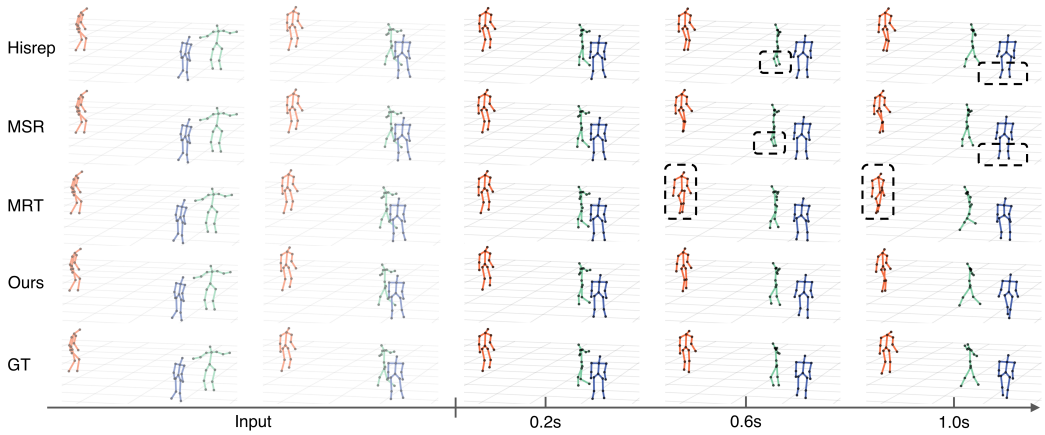

Figure 5. Qualitative comparison with the baselines and the ground truth on a sample of the CMU-Mocap (UMPM) dataset. The left two columns are inputs, and the right three columns are predictions.

the people order of input data in the model should not affect its performance, we randomly permute this order during training and testing to investigate model robustness with the results in Tab. 4. Obviously, our method is just as robust as the single person-based methods, *i.e.*, do not rely on permutation of person in the input.

4.7. Attention Visualization

We show the visualization of attention score between individuals' query motion and the historical context of different people in Fig. 6 (b). The left figure shows the observed motion of three people, where we can see that person 3 (P_3) is following person 2 (P_2) around, while person 1 (P_1) is not interacting with them nearly. The right figure draws the corresponding attention score for each individual. High attention scores for the two individuals interacting are indicated by two red-dotted regions. In terms of the high interaction group, in practice, P_3 should pay more attention to historical information about P_2 in order to adjust his behavior, which is clearly demonstrated through the visualization.

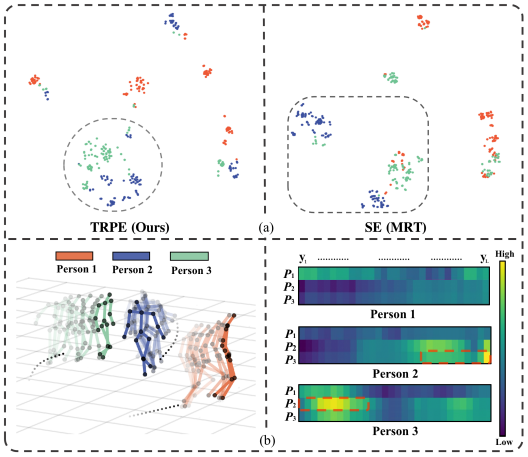

Figure 6. (a) Comparative visualization of feature distributions in t-SNE representations. The left figure shows the results obtained from our model equipped with TRPE, while the middle figure shows the results obtained from the MRT model with Spatial Encoding (SE). (b) Attention visualization of the first layer in Transform decoder. The x-axis denotes the input sequence from timestamp 1 to L, and the y-axis represents different individuals.

5. Conclusion

In this paper, we presented a novel Transformer architecture for effective multi-person pose forecasting. We first constructed a TBPM to extract spatial and temporal features based on body semantics. We also presented an SBI-MSA module to learn body part dynamics for inter- and intra-individual interactions. In addition, we proposed a novel Trajectory-Aware Relative Position Encoding for SBI-MSA to offer discriminative spatial information and additional interactive clues. Experiments demonstrated that our method outperformed state-of-the-art methods on multiple motion datasets.

Limitations and Social Impacts. Our work does not come without limitations. MPBP sequence involves all the individuals' body parts and time information. When inputting a long series containing many people, it will lead to heavy attentional computation during training and inference. We plan to address this issue in future. For social impacts, we are still uncertain as to whether a person can be identified based purely on his or her poses and movements. However, compared to input images of people, it is harder to invade individuals' private information.

Acknowledgments. This work was partially supported by the Zhejiang Provincial Natural Science Foundation (LGF21F20012) and Zhejiang Provincial Science and Technology Program in China (2021C03137).

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 2

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 1, 2

[3] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 3, 5

[4] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2(3):5, 2020. 1

[5] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. *computer vision and pattern recognition*, 2016. 3

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 1

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 1, 2

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 1

[9] CMU-Graphics-Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003. 2, 6

[10] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4801–4810, 2021. 2

[11] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017. 5

[12] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. 2, 6, 7

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 1, 2

[14] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the european conference on computer vision (ECCV)*, pages 786–803, 2018. 1

[15] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 1, 2

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[17] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281, 2019. 3

[18] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[19] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard S. Zemel. Neural relational inference for interacting systems. *international conference on machine learning*, 2018. 3

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Learning*, 2022. 3

[21] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019. 1

[22] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *neural information processing systems*, 2020. 3

[23] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. *ArXiv*, abs/2208.00368, 2022. 2

[24] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 1

[25] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Transactions on Image Processing*, 30:7760–7775, 2021. 1, 2

[26] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2, 6, 7

[27] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 3

[28] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 1

[29] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2

[30] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2, 6

[31] Theodore M Newcomb, Ralph H Turner, and Philip E Converse. *Social psychology: The study of human interaction*. Psychology Press, 2015. 1

[32] Hiroaki Sakoe. Dynamic-programming approach to continuous speech recognition. In *1971 Proc. the International Congress of Acoustics, Budapest*, 1971. 4

[33] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978. 4

[34] Pengxiang Su, Zhenguang Liu, Shuang Wu, Lei Zhu, Yifang Yin, and Xuanjing Shen. Motion prediction via joint dependency modeling in phase space. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 713–721, 2021. 1, 3

[35] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 1264–1269. IEEE, 2011. 2, 6

[36] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 7

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4, 5

[38] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2, 6

[39] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 1, 2, 3, 6, 7

[40] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021. 5

[41] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022. 3

[42] Yi Xu, Lichen Wang, Yizhou Wang, and Yun Fu. Adaptive trajectory prediction via transferable gnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6520–6531, 2022. 3

[43] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. *european conference on computer vision*, 2020. 3

[44] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Unsupervised online video object segmentation with motion property understanding. *IEEE Transactions on Image Processing*, 29:237–249, 2019. 1