幻灯片 1

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

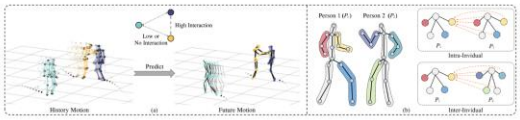
用于多人姿势预测的轨迹感知身体交互转换器，这个转换器命名为TBIFormer。

多人姿势预测目前仍然是一个具有挑战性的问题，尤其是在复杂人群场景中，对人体交互进行建模时。现有的姿势预测方法通常将整个姿势序列表示为时间序列，但忽略了人与人之间的交互影响。文章因此提出了一种全新的轨迹感知身体交互转换器（TBIFormer），通过建模身体部位的交互来进行多人姿势预测。

幻灯片 2

Multi-Person Pose Forecasting

- Current work
  - RNNs, Transformers; GNNs;
  - based on local pose dynamics
  - without considering global position changes of body joints (global body trajectory)
  - overlooking human interaction
- Ours
  - a novel Transformer-based framework——**TBIFormer**



行人轨迹预测是多人社会交互　的一个代表性问题。现有的任务方法可以根据它们如何建模时间和社会维度进行分类。 RNNs [16] 和 Transformers [37] 是处理时间建模轨迹序列的首选模型 [5, 17, 43]，图神经网络 (GNNs) [20] 通常被用作交互建模的社会模型 [19 ,22,41,42]。虽然表现良好，但这些方法大多基于局部姿态动力学预测，而没有考虑身体关节的全局位置变化（全局身体轨迹）。实际上，在现实世界的场景中，每个人都可能与一个或多个人进行交互，交互程度从低到高，具有即时和延迟的相互影响 [2, 31]。并且往往在忽视人与人之间的相互作用，只是解决孤立的单个人的问题

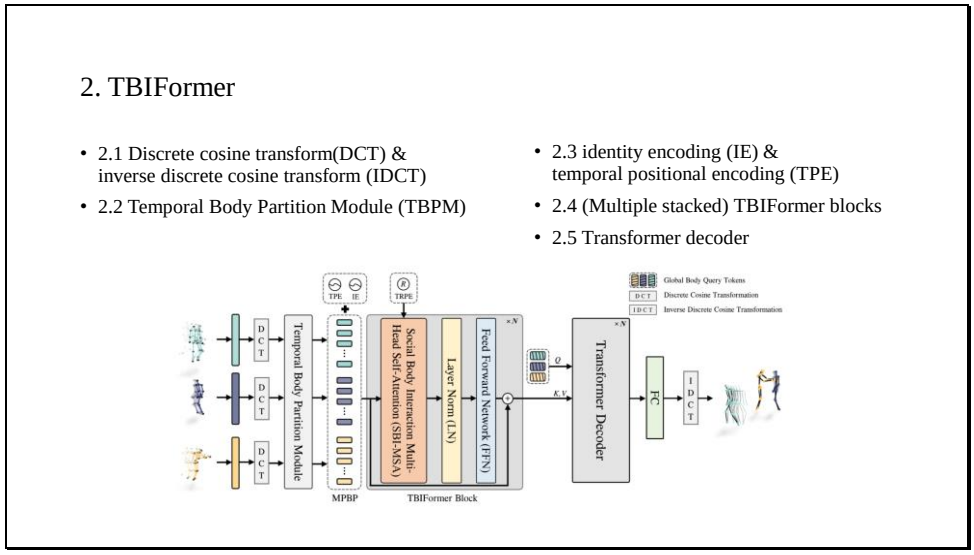
实际上，在现实世界的场景中，每个人都可能与一个或多个人进行交互，交互程度从低到高，具有即时和延迟的相互影响 [2, 31]。如图 1 (a) 所示，两个人在高度互动的情况下推搡，而第三个人在没有互动或互动很少的情况下漫步。因此，准确预测姿势动力学和轨迹并综合考虑复杂的社会交互因素对于理解多

人运动预测中的人类行为势在必行。然而，现有的解决方案并不能有效地解决这些具有挑战性的因素。

文章提出的 TBIFormer，通过对个体之间的骨骼身体部位，进行动力学建模，来考虑细粒度的人体交互，并预测 3 ~ 10 人在 3D 场景中的未来运动。

文章将身体关节分为 5 个部分，考虑两种互动。一种是Intra-Individual 分支用于探索每个个体内部的部分关系，Inter-Individual 分支旨在捕获不同个体之间身体部位的交互依赖性。也就是说 TBIFomer 有助于同时模拟个体内和个体间的身体部位相互作用。

幻灯片 3

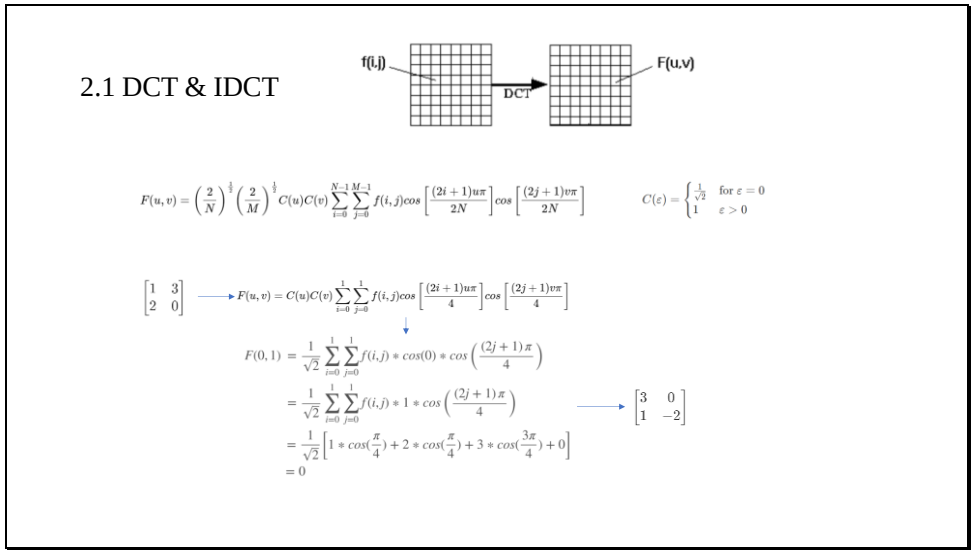


TBIFormer整体框架

它包含多个堆叠的 TBIFormer 块和一个transformer decoder转换器解码器，后跟全连接层。

给定观察到的 3 个人的姿势序列，TBIFormer 将它们转换为位移序列作为输入，然后预测每个人的未来姿势。在 TBIFormer 的头部和尾部，我们采用离散余弦变换 (DCT) [3]丢弃高频信息以在位移轨迹空间中实现更紧凑的表示。接下来是一个时间身体分区模块TBPM，将所有姿势序列转换为多人 BodyPart 序列，旨在更好地学习和保留骨骼序列中身体部位的空间和时间信息。此外，还引入了时间位置编码TPE、人员身份编码IE和轨迹感知相对位置编码TRPE，以保存时间、身份和可区分的空间信息。每个 TBIFormer 块都有一个social body interaction multi-head self-attention(SBI-MSA)社交身体交互多头自体- 注意力模块，用于对跨时间和社会维度的身体部位交互进行建模。

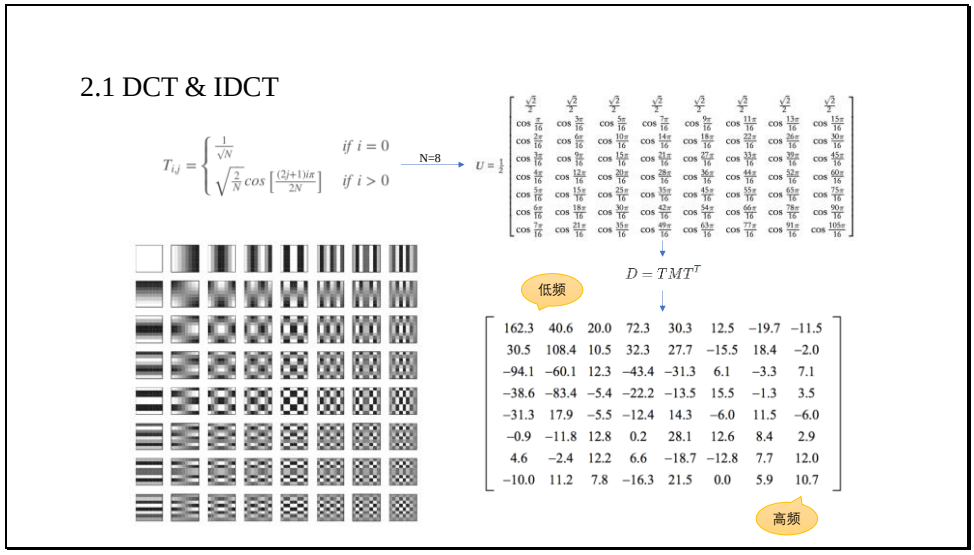
幻灯片 4



DCT全称是离散余弦变换。

DCT 由如下的公式定义，可以看出将图像像素从空间域转换到频域。N 和 M 为矩阵的行数和列数，u,v = 离散频率变量(0,1,2¡-7)，f(i,j) = 图像在 i 行 j 列的像素值。以一个简单地二维矩阵为例，将其像素逐个带入到公式中。要做4次运算，这里放出了(0,1)像素的具体计算，四次运算后最终得到了DCT矩阵。

幻灯片 5



在实践上，上述方式的计算效率不高，更加简便的计算方式是使用 DCT 矩阵，如果N取8可以得到右边的DCT矩阵。

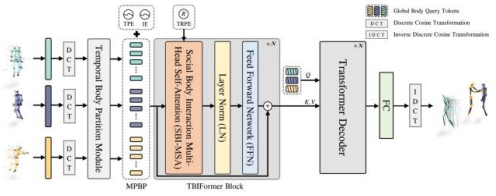
这个矩阵最左上角是直流分量，往右下角逐渐频率变高，从低频信息，一直到最右下角的高频信息。

有一种东西叫基准样式，左下角。对于任意一张 8*8 的图，都可以由标准图中的的小图以一定比例叠加而合成，而每个小图的权重，由 DCT 变换得到的矩阵决定。DCT 变换后左上角一般是比较大的数字，而右下角一般是比较小的数字，也就是说图片中低频信息占的比重较多，DCT一般在做图片或者视频编码压缩的时候，就是通过量化舍弃右下角的高频信息，来实现信息的压缩。

幻灯片 6

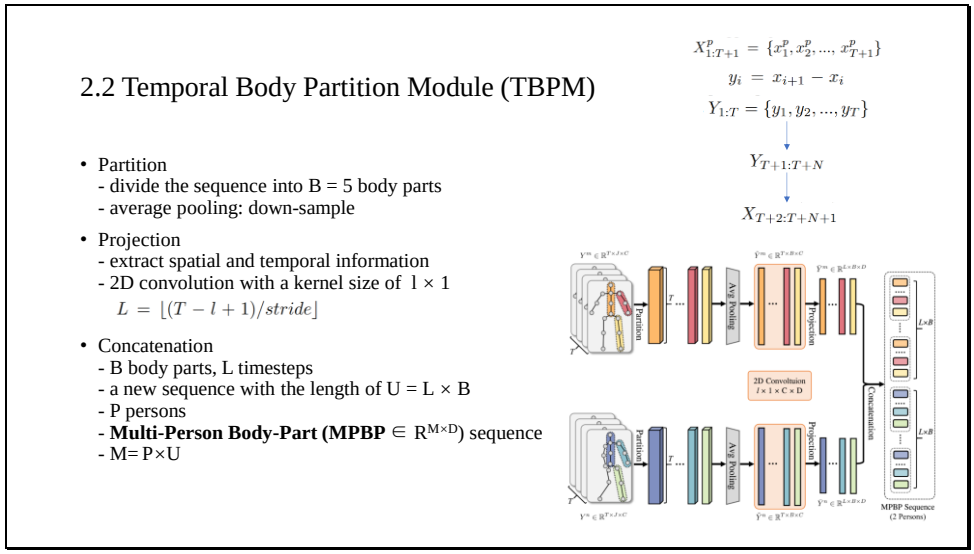
2. TBIFormer

- 2.1 Discrete cosine transform(DCT) & inverse discrete cosine transform (IDCT)
- 2.2 Temporal Body Partition Module (TBPM)
- 2.3 identity encoding (IE) & temporal positional encoding (TPE)
- 2.4 (Multiple stacked) TBIFormer blocks
- 2.5 Transformer decoder



为了向 TBIFormer 提供包含时间和空间信息的姿势序列，一种直观的方法是在时间序列中保留身体关节。也就是直接使用原始的关节序列，但是现实中可能因为嘈杂的传感器输入，或者不准确的估计，引起的嘈杂关节而使得结果不准确。在这项工作中，我们提出了一种基于人体语义的时间身体分割模块(TBPM)，不直接采用原始的姿势序列，将原始姿势序列转换为新姿势序列，从而增强网络对交互式身体部位建模的能力。这就是TPBM做的工作。

幻灯片 7



假设对第p个人观察到的骨骼序列是X集合，有T+1帧的数据。本文不用世界坐标中的绝对关节位置，而是用y来表示第i时刻的瞬时位姿位移，这种方法能够更好的涵盖动力学信息。给定Y，从1到T帧，我们的目标是得到T+1帧和T+N帧，再转换到世界坐标中得到XT+2到T+N+1。

分为三部分。分区、投影和连接。

- 分区：给定第p个人的位移序列 Y p ∈ RT ×J×C ，其中T和J分别表示帧数和关节数，C = 3表示3D坐标的维度，我们首先将序列进行划分，基于自然人体骨骼结构将序列分成B = 5个身体部位（例如左右臂，左右腿，核心躯干），然后通过平均池化对每个身体部位进行下采样。经过上述操作，序列表示为 ˜ Y p ∈ RT ×B×C 。维度发生变化，从原来J个关节降维到了B个关节组。

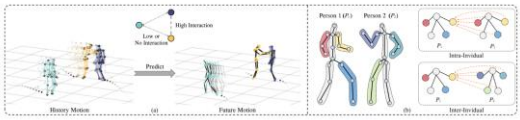
- 投影。投影操作的作用是初步提取空间和时间信息。具体来说，在 ˜ Y p 上使用核大小为 I × 1 的 2D 卷积，可以得到 2D 特征图 Y p ∈ RL×B×D，其中 L = ⌊(T − I + 1)/stride⌋，D 表示输出通道数。
- 投影之后，所有 B 个身体部位的编码在所有 L 个时间步中被连接起来，形成一个长度为 U = L × B 的新序列。接下来，我们将所有 P 个人的序列一个接一个地连接起来，形成一个合并的多人身体部位 (MPBP ∈ RM×D) 序列，其中 M 表示 P × U。 MPBP 序列允许我们的 TBIFormer 跨时间和社会维度学习个体的身体部位动态

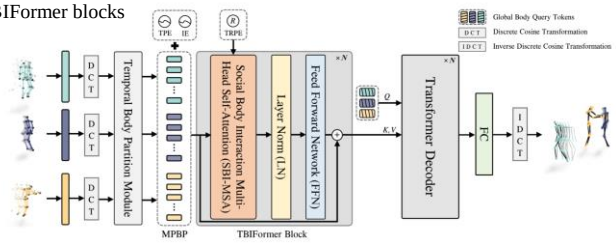
通过以上的序列转换，TBPM能够更好地保留骨架序列的空间和时间信息，用于后续transformer。

幻灯片 8

2. TBIFormer

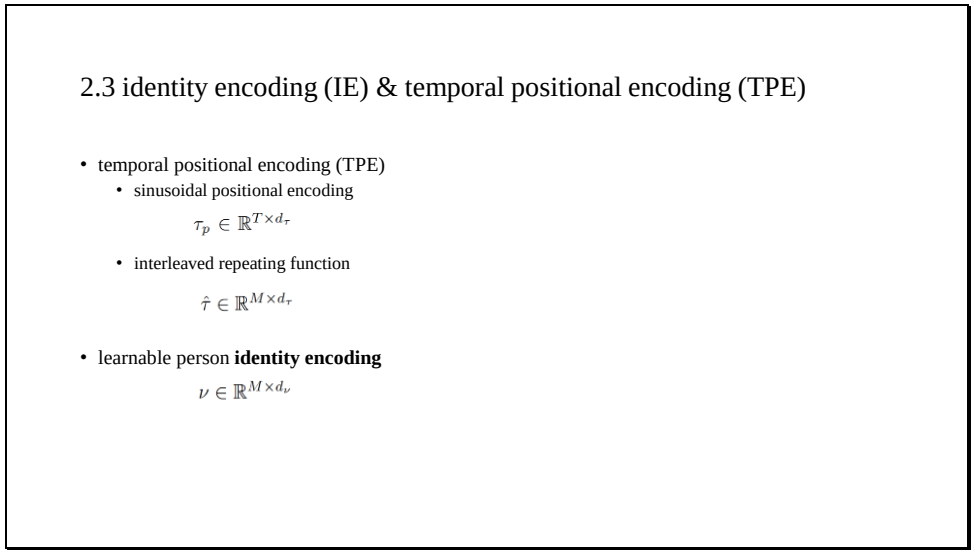
- 2.1 Discrete cosine transform(DCT) & inverse discrete cosine transform (IDCT)
- 2.2 Temporal Body Partition Module (TBPM)
- 2.3 identity encoding (IE) & temporal positional encoding (TPE)
- 2.4 (Multiple stacked) TBIFormer blocks
- 2.5 Transformer decoder





类似于Transformer [37]，文章应用正弦位置编码来向 TBIFormer 传达与 MPBP 序列中每个元素关联的时间步长。

幻灯片 9

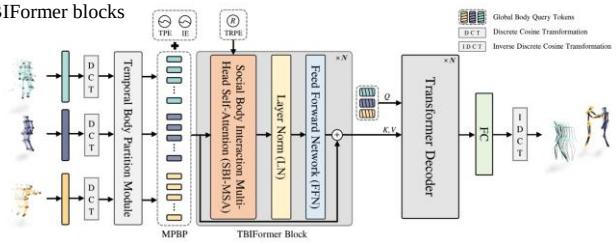


我们首先根据每个人的时间步计算时间戳特征，而不是根据整个 MPBP 序列中的索引对每个元素的位置进行编码，并获得时间位置编码 τp ∈ RT ×dτ，其中 dτ 是时间戳的特征维度。然后我们利用交错重复函数重复 B 个身体部位的编码元素，并连接所有个体的编码。最终的时间位置编码 (TPE) 公式为 ^ τ ∈ RM×dτ 。

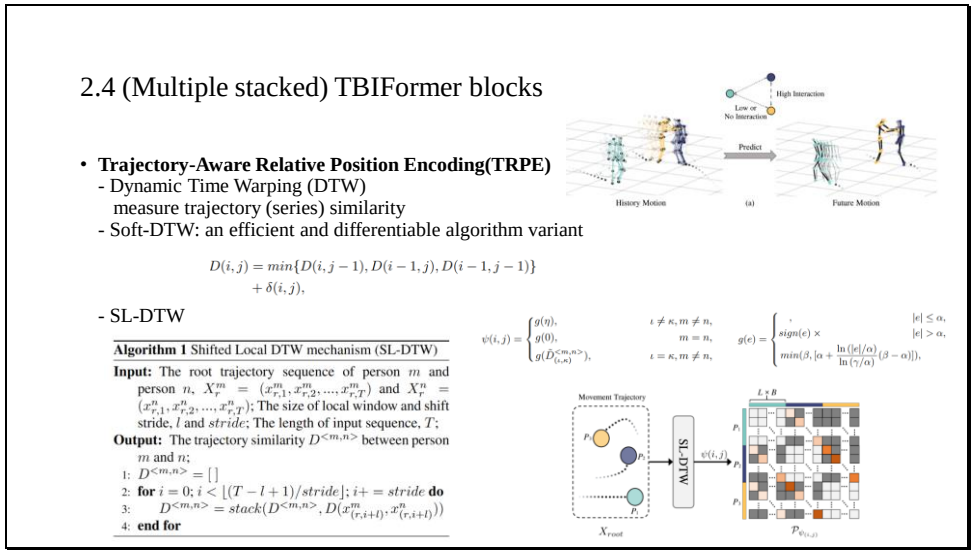
为了提供 MPBP 序列中每个个体的身份信息，我们还注入了一个可学习的人身份编码 ν ∈ RM×dν ，指示每个元素属于哪个个体，其中 dν 表示特征维度。值得注意的是，身份编码 (IE) 是随机初始化的，并使用与 TPE 相同的重复方法重复时间和身体部位。

2. TBIFormer

- 2.1 Discrete cosine transform(DCT) & inverse discrete cosine transform (IDCT)
- 2.2 Temporal Body Partition Module (TBPM)
- 2.3 identity encoding (IE) & temporal positional encoding (TPE)
- 2.4 (Multiple stacked) TBIFormer blocks
- 2.5 Transformer decoder



幻灯片 11



TRPE模块

一个本能的假设是，几个人之间的距离越近，他们之间的互动可能就越高。然而，在复杂的人群情况下，一个人可能会背对着附近的一个人而没有互动，或者如图 1 所示，一个人可能只是从两个互动的人身边经过，表现出低互动但很接近。因此，基于欧几里得距离的空间位置编码很难提供有区别的空间信息并区分实际交互的个体。

在本文中，我们的观察是，在 3D 空间中交互的人倾向于沿相同或面对面的方向移动，而不是偏离方向。主要挑战是直接计算人体骨骼数据的身体方向和个体之间的角度是繁琐且昂贵的。为了解决这个问题，我们发现运动轨迹也可以提供重要信息并规避上述限制。因此，我们通过测量运动轨迹的相似性提出了一种新的轨迹感知相对位置编码（TRPE），它可以聚合相应的运动模式和空间信息。动态时间规整（DTW）[32, 33] 是一种比欧氏距离更稳健的测量轨迹（系

列）相似性的方法。在这项工作中，我们采用了一种高效且可微分的算法变体，称为 Soft-DTW [11]，它可以定义为，

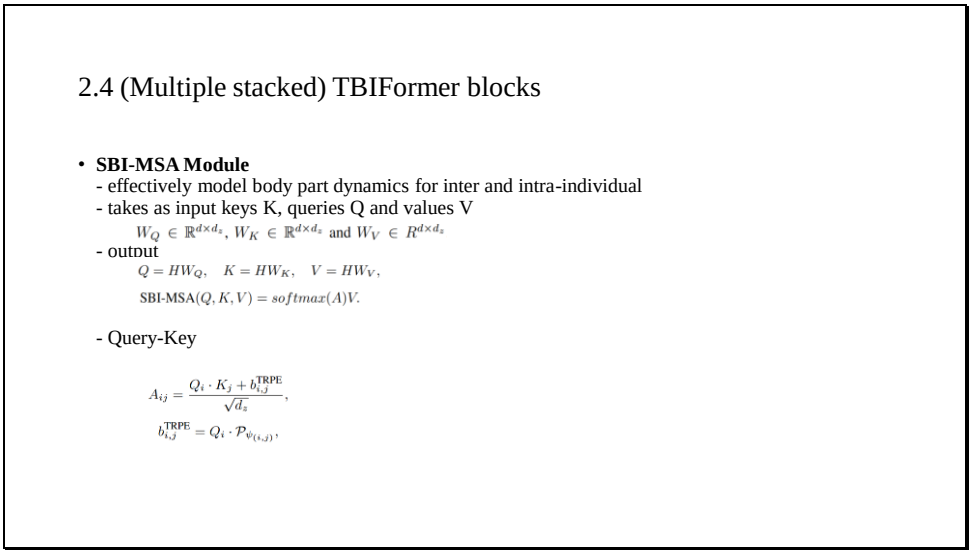
其中 D(i, j) 表示子序列 S1 = (s1, s2, ..., si) 和 S2 = (s1, s2, ..., sj) 之间的最短距离，δ(·,·) 是可微分成本功能。为了根据某个时间步而不是完整的时间戳动态获取轨迹相似性，我们提出了一种基于 SoftDTW 的移位局部 DTW (SL-DTW) 机制。与卷积运算类似，SL-DTW在特定的窗口大小下计算个体之间的相似度并逐步移动，这将提供更精确的相关信息。有关 SL-DTW 过程的详细描述

给定 P 个人之间的轨迹相似度距离 ˜ D ∈ RP ×L，我们需要将距离映射到一个整数集以进行相对位置编码。解决这个问题的常用方法是裁剪函数：h(˜ D) = max(−β, min(β, ˜ D))，这不可避免地消除了远距离相对位置的上下文。因此，我们替代地使用分段函数 [40] g(·) 维护长程信息以将相对距离索引到相应的编码，然后通过 SL-DTW 距离定义索引矩阵如下：

其中 [·] 是一个舍入操作，sign() 确定数字的符号，即返回 1 为正输入，-1 为负，否则为 0。 α 控制分段点，β 将输出限制在 [−β, β] 范围内，γ 调节对数部分的曲率。

最后，如图 4 所示，我们将轨迹相似度的索引矩阵 ψ(i, j) 嵌入为我们的 TRPE Pψ(i,j) ∈ RM×dz 其中，M = P × L × B，在 SBI-MSA 的所有注意力层中共享。

幻灯片 12



SBI-MSA Module

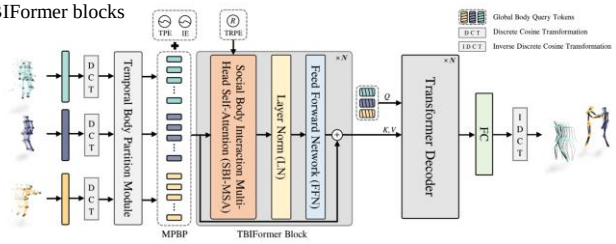
在每个 TBIFormer 块中，我们的目标是构建一个社交身体交互多头自注意力 (SBI-MSA) 模块，以有效地模拟个体间和个体内的身体部位动态。鉴于 TBPM 提取的运动特征，基于运动注意计算的 SBI-MSA 可以进一步优化姿势动力学并捕获个体之间复杂的身体交互依赖性。令 H = [h1, ..., hn] ∈ Rn×d 表示注意力模块的输入表示，其中 d 是隐藏维度。 SBI-MSA 将输入键 K、查询 Q 和值 V 作为输入，每个值都由相应的参数矩阵 WQ ∈ Rd×dz、WK ∈ Rd×dz 和 WV ∈ Rd×dz 投影。 SBI-MSA 的输出计算为

我们将 TRPE Pψ(i,j) 整合到注意力图上，以考虑个体动态特征与跨时间和社会维度的空间线索之间的相互作用。将 Aij 表示为 Query-Key 乘积矩阵 A 的 (i, j) 元素，我们有其中 bTRPE i,j 是注意力图的上下文偏差。

幻灯片 13

2. TBIFormer

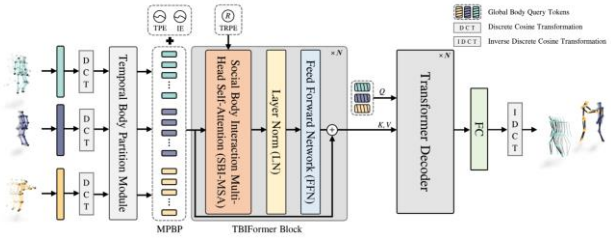
- 2.1 Discrete cosine transform(DCT) & inverse discrete cosine transform (IDCT)
- 2.2 Temporal Body Partition Module (TBPM)
- 2.3 identity encoding (IE) & temporal positional encoding (TPE)
- 2.4 (Multiple stacked) TBIFormer blocks
- 2.5 Transformer decoder



幻灯片 14

2.5 Transformer decoder

- concatenate joint coordinates of the last observed sub-sequence (length = l)
- down-sample them on time dimension by 1D Convolution (kernel size = l)



如图 2 所示，我们将来自每个人的所有身体关节的最后观察到的子序列（长度 = l）的关节坐标连接起来，并通过一维卷积（内核大小 = l）在时间维度上对它们进行下采样作为全局正文查询标记。键和值标记是 TBIFormer 块的输出。我们利用标准的 Transformer 解码器 [37] 来编码当前（查询）和历史上下文（键）之间的关系。在解码器的最后，我们采用两个完全连接 (FC) 层，然后是逆离散余弦变换 (IDCT) [3] 来生成每个个体的未来运动轨迹 XT +2:T +N+1。

幻灯片 15

3. Experiment

- 数据集
  - CMU-Mocap , CMU-Mocap (UMPM), 3DPW, MuPoTs-3D
  - 训练： CMU-Mocap (UMPM)
  - 测试：CMU-Mocap (UMPM) , MuPoTs-3D（2 ~ 3 人) , Mix1 （6 人） & Mix2 （10 人）
- Evaluation
  - JPE Metric
  - APE Metric
  - FDE Metric

$$\mathrm{JPE}(X,\hat{X}) = \frac{1}{P \times J}\sum_{i=1}^{P}\sum_{j=1}^{J}||X_j^i - \hat{X}_j^i||^2, \qquad \mathrm{AME}(X,\hat{X}) = \mathrm{JPE}(X - X_r, \hat{X} - \hat{X}_r),$$

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

- Baselines
  HRI：基于注意力
  MSR：基于GCN
  MRT

为了验证 TBIFormer 的有效性，我们在 CMU-Mocap (UMPM) 数据集上运行实验，它将 UMPM [35] 合并到 CMU-Mocap [9] 中以进行数据集扩展。 Mix1 和 Mix2 由 CMU-Mocap、UMPM、3DPW [38] 和 MuPoTs-3D [30] 数据集混合。我们通过在 MuPoTS-3D（2 ~ 3 人)、Mix1（6 人）和 Mix2（10 人）数据集上进行测试来评估所有方法的泛化能力，模型仅在 CMU-Mocap (UMPM) 数据集上训练。关于为什么要做数据集扩展以及混合数据集的处理细节，请参阅附录。

JPE：我们使用基于平均每个关节位置误差 (MPJPE) 的关节位置误差 (JPE) 来测量所有个体的姿势，包括身体轨迹：

APE：我们移除全局运动并使用对齐平均每个关节位置误差 (APE) 来测量纯位姿位置误差：其中 Xr 和 ^ Xr 是人体的估计根位置和真实根位置。

FDE：我们还采用根位置来评估每个人的全局运动，使用典型的轨迹预测指

标：最终位移误差 (FDE)。公式说明如下：其中 Xr,N 和 ^ Xr,N 是最终姿态在第 N 个预测时间戳处的估计和真实根位置。

我们选择 3 种代码发布的最先进 (SOTA) 方法作为基线，包括两种基于单人的方法：HRI [26] 和 MSR [12]，以及最近发布的基于多人的方法 MRT [39]。HRI [26] 是一个基于注意力的网络，MSR [12] 是一个基于 GCN 的方法，它们都允许绝对坐标作为输入。对于短期预测，我们使用 50 帧（2.0s）的输入和 25 帧（1.0s）的预测来训练所有这些模型，并在 4 个数据集上进行评估。对于长期预测，使用 MRT [39] 中的协议，我们设置 15 帧（1.0s）的历史作为输入来预测未来的 45 帧（3.0s）。

幻灯片 16

Results

		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**

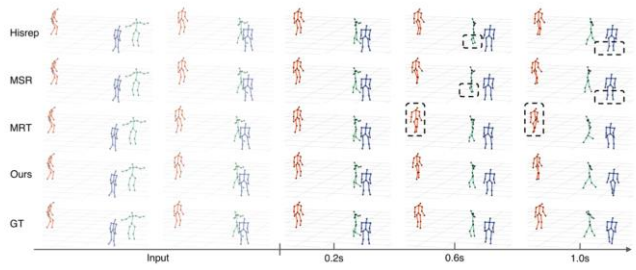
为了验证 TBIFormer 的预测性能，我们遵循大多数单人方法 [12, 26] 的设置来显示短期和长期预测的定量和定性结果，并将我们的方法与基线进行比较。

定量结果。表 1 报告了 JPE、APE 和 FDE 在 4 个不同数据集上的结果。我们的 TBIFormer 在预测准确性方面明显优于基线。与之前基于单人的方法相比，我们实现了高达 13% ~ 27% 的改进，与基于多人的方法相比，实现了高达 13% ~ 16% 的改进。可以注意到，由于缺乏人体骨骼的空间建模，MRT [39] 在 APE 指标中表现不佳。此外，我们在表中报告了长期预测（1.0s~3.0s）的结果。 2. 我们的方法在 3 个指标中始终优于基线。

[JPE、APE 和 FDE（以毫米为单位）在不同数据集上的结果。我们将我们的方法与之前用于短期和长期预测的 SOTA 方法进行了比较。最佳结果以粗体显示。 （*表示多人运动预测方法。）]

幻灯片 17

Results



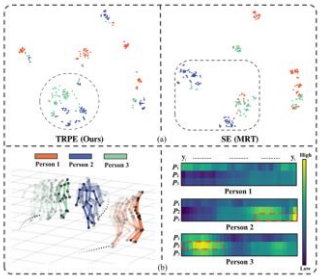
与CMU-Mocap (UMPM) 数据集，定性比较。左边两列是输入，右边三列是预测。

HRI [26] 和 MSR [12] 的结果表明，它们在长期预测中倾向于收敛到静态姿势。由于人体空间建模的不足，MRT [39] 产生了一些扭曲的姿势。相比之下，我们的方法在实践中生成了更合理的 3D 人体运动，这比其他方法更接近地面实况。

幻灯片 18

Ablation studies

	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 → 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**



TBPM 的有效性。 TBPM 构建了一个序列，其中包含人体姿势的时间和空间信息。当它被移除并且关节坐标直接连接到姿势序列中的身体关节时，TBIFormer 无法学习身体部位动力学，我们可以观察到性能显着下降。

IE 和 TRPE 的有效性。 Person identity encoding (IE) 允许我们的方法区分 MPBP 序列中的元素类型（即，通知每个标记有关身份信息）。去除 IE 后，模型的整体性能略有下降。轨迹感知相对位置编码 (TRPE) 为模型提供了充足的空间和交互线索。当我们移除 TRPE 时，性能会大幅下降。此外，如表中的 (TRPE → EuPE) 所示。 3，即使用基于欧氏距离的位置编码替换 TRPE 后，性能仍然不是最优的。我们还提供 t-SNE 可视化 [36] 以证明 MRT [39] 中 TRPE 和 SE（基于欧几里德的编码）之间的辨别力。显然，我们配备 TRPE 的模型可以获得更准确和紧凑的表示。

SBI-MSA 的有效性。 SBI-MSA 的目标是跨时间和社会维度学习身体部位的动态。如选项卡的最后一行所示。 3，如果将 SBI-MSA 替换为标准的自注意力模块，我们的模型只会单独学习每个人的运动特征，导致长期性能较差。

随机人排列的影响。为了保证模型中输入数据的人员顺序不应影响其性能，我们在训练和测试期间随机排列此顺序，以使用表中的结果研究模型的鲁棒性。4. 显然，我们的方法与基于单人的方法一样稳健，即不依赖于输入中人的排列。

右图展示了个人查询动作与不同人的历史背景之间注意力得分的可视化。左图显示了观察到的三个人的动作，我们可以看到人 3 (P3) 正在关注人 2 (P2) 周围，而人 1 (P1) 几乎没有与他们互动。右图绘制了每个个体对应的attention score。两个红点区域表示两个人交互的高注意力分数。对于高交互组，在实践中，P3应该更多地关注P2的历史信息，以调整他的行为，这一点通过可视化表现得很清楚

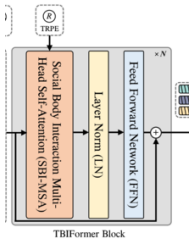
t-SNE 表示中特征分布的比较可视化。左图显示了我们配备 TRPE 的模型获得的结果，而中间图显示了使用空间编码 (SE) 的 MRT 模型获得的结果。 (b) Transform 解码器第一层的注意力可视化。 x轴表示从时间戳1到L的输入序列，y轴表示不同的个体。

幻灯片 19

Conclusion

- A novel **Transformer architecture** for effective multi-person pose forecasting
  - Temporal Body Partition Module (TBPM)
  - Social Body Interaction Multi-Head Self-Attention (SBI-MSA)|
  - Trajectory-Aware Relative Position Encoding(TRPE)

- Limitations
  - heavy attentional computation during training and inference



1) 我们提出了一种新颖的基于 Transformer 的有效多人姿势预测框架，并设计了一个时间身体分区模块，将原始姿势序列转换为多人身体部位序列以保留时间和空间信息。 2) 我们提出了一种新颖的社会身体交互多头自我注意 (SBI-MSA)，它可以跨个体间和个体内部学习身体部位动态并捕获复杂的交互依赖性。 3) 为 SBI-MSA 引入了一种新的轨迹感知相对位置编码，以提供有区别的空间信息和额外的交互线索。

Our work does not come without limitations. MPBP 序列涉及所有个体的身体部位和时间信息。当输入包含很多人的长序列时，会导致在训练和推理过程中进行大量的注意力计算。