

Efficient Multi-Person Motion Prediction by Lightweight Spatial and Temporal Interactions

Yuanhong Zheng¹ Ruixuan Yu^{1*} Jian Sun^{2,3}

¹School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China

²School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

³Pazhou Laboratory (Huangpu), Guangzhou, Guangdong 510555, China

202200800358@mail.sdu.edu.cn, yuruixuan@sdu.edu.cn, jiansun@xjtu.edu.cn

Abstract

3D multi-person motion prediction is a highly complex task, primarily due to the dependencies on both individual past movements and the interactions between agents. Moreover, effectively modeling these interactions often incurs substantial computational costs. In this work, we propose a computationally efficient model for multi-person motion prediction by simplifying spatial and temporal interactions. Our approach begins with the design of lightweight dual branches that learn local and global representations for individual and multiple persons separately. Additionally, we introduce a novel cross-level interaction block to integrate the spatial and temporal representations from both branches. To further enhance interaction modeling, we explicitly incorporate the spatial inter-person distance embedding. With above efficient temporal and spatial design, we achieve state-of-the-art performance for multiple metrics on standard datasets of CMU-Mocap, MuPoTS-3D, and 3DPW, while significantly reducing the computational cost. Code is available at <https://github.com/Yuanhong-Zheng/EMPMP>.

1. Introduction

3D multi-person motion prediction focuses on predicting the future positions of skeletal joints for multiple individuals based on the historical motions. It has broad applicability across various domains, including human-computer interaction [5, 25, 34, 35], virtual/augmented reality [13, 16, 36, 37], sports analysis [6, 24], and surveillance systems [8, 18]. This task is challenging as it necessitates not only accurate prediction of individual poses through the modeling of historical temporal data, but also a comprehensive representation of inter-person interactions.

Early studies for 3D motion prediction commonly focused on single-person pose forecasting [15, 22, 27, 28,

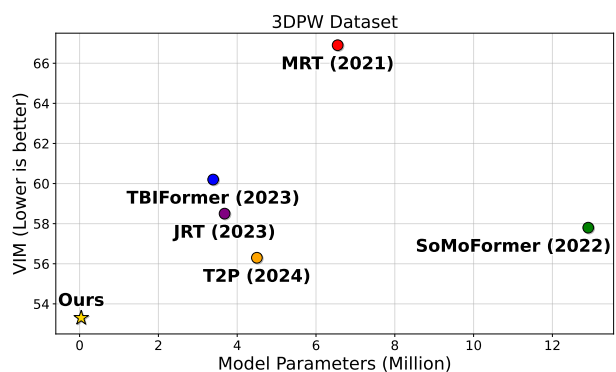


Figure 1. Comparison of our model with state-of-the-art methods on the 3DPW dataset. Our approach outperforms others while requiring significantly fewer parameters.

30, 31, 39, 46], where the graph neural networks and Transformer-based networks were commonly employed to extract joint-wise features. For example, PGBIG [28] proposed to alternately extract features through spatial and temporal graph convolutional networks, while STCFormer [39] learned cross-joint and cross-frame interaction with Spatial-Temporal Criss-cross attention. Though effective for individual poses, these methods ignored inter-person interactions critical for multi-agent scenarios. Meanwhile, several works on multi-agent trajectory prediction [4, 10, 21, 49] have been proposed by treating the agents as points within an interactive system, and conducting trajectory prediction with local context extraction and global interaction modeling. However, these works overlook the intrinsic structural information of individual agents, such as the human skeletal configuration.

Recent research has increasingly focused on multi-person motion prediction, integrating trajectory estimation for global movement and pose forecasting for fine-grained joint-level dynamics [20, 33, 41, 45, 47]. MRT [45] proposed a multi-scale Transformer model with two encoders to capture individual and inter-person interactions respec-

*Corresponding author

tively. TBIFormer [33] conducted motion prediction by modeling the body part interactions. JRT [47] employed Transformer-based architectures to model both individual joints and inter-person relation, while T2P [20] separately modeled human pose and trajectory. To the best of our knowledge, all of these methods rely on Transformer-based architectures to capture spatial or temporal interactions. Despite their success in multi-person motion prediction tasks, these models often encounter challenges, including the large number of parameters and the high computational cost associated with Transformer architectures. Furthermore, while they predominantly emphasize the spatial information, there has been limited exploration on the temporal sequential data.

In this work, we propose the Efficient Multi-Person Motion Prediction (EMPMP) network, which leverages a lightweight architecture to effectively capture both spatial and temporal interactions. We first introduce the Multi-level Estimation (ME) block, designed to efficiently extract spatial and temporal features for both individual and inter-person representations. Furthermore, we present the Cross-level Interaction (CI) block, which facilitates the fusion and updating of multi-level features through learned affine transformations. To enhance the modeling of social interactions, we explicitly incorporate inter-person distance into individual representation. Extensive experiments on standard datasets demonstrate that our approach achieves state-of-the-art performance across multiple evaluation metrics, while utilizing only 1% to 10% of the parameters compared to existing methods. As depicted in Fig. 1, our model outperforms others on the 3DPW dataset, achieving superior performance with a minimal number of parameters.

2. Related Works

Single-Person Pose Forecasting. The task of single-person pose forecasting aims to predict future human poses based on past observations. Early approaches primarily relied on models such as Hidden Markov Models (HMM) [22] and Gaussian Processes (GP) [44]. Recurrent Neural Networks (RNNs) were also widely applied for sequential human pose forecasting [12, 15, 27, 43]. For example, ERD [15] combined nonlinear encoders and decoders with RNNs to jointly model human representation and dynamics, while NAFMP [27] utilized Lie algebra and a hierarchical recursive network to capture both local and global contexts. To mitigate error accumulation in frame-by-frame predictions, feed-forward models like Graph Convolutional Networks (GCNs) and Transformer-based architectures were introduced to model long-range dependencies [14, 28, 30, 31]. For instance, LTD [30] proposed learning graph connectivity to capture long-range dependencies in joint sequences, while HRI [31] leveraged attention mechanisms to guide predictions based on historical patterns. Despite their suc-

cess in pose forecasting tasks, these methods are mainly restricted to single-person scenarios, and they are not suitable for real-world applications involving multiple individuals and inter-person interactions.

Multi-Person Motion Prediction. This task focuses on predicting the individual movements of people while capturing their complex interactions with others in the scene. Current approaches commonly rely on Transformer-based networks [20, 33, 41, 45, 47], with various design strategies to capture spatial or temporal information. For example, MRT [45] combines local and global encoders to capture both individual and multi-person features, while TBIFormer [33] transforms pose sequences into body-semantic-based part sequences for fine-grained interaction modeling. JRT [47] models joint relationships using relative distances and physical constraints. SoMoFormer [41] represents motion as joint sequences and encodes joint types, identities, and global positions. T2P [20] decouples the task into local pose and global trajectory predictions. While these methods effectively model interactions, they often require high parameter counts and computational resources. In this work, we propose a novel lightweight network to efficiently captures spatial and temporal representations as well as their interactions, achieving state-of-the-art results with significantly lower computational cost.

MLP-based Lightweight Neural Networks. It is crucial to maintain a lightweight architecture while ensuring high performance. Conventional approaches, such as model pruning [38], knowledge distillation [26], and sparse attention [11], have been widely adopted. MetaFormer [48] showed that the Transformer’s strength primarily lies in its overall design rather than Self-Attention, motivating MLP-based alternatives. MLP-Mixer [40] and CycleMLP [9] replace attention with linear layers, achieving competitive performance with fewer parameters. For the single-person 3D pose forecasting task, Motion Mixer [7] captures spatial and temporal dependencies with separate MLPs, while SiMLPe [17] integrates DCT [3] and cascaded MLPs to strengthen temporal representations. GraphMLP [23] combines MLPs and GCN in a unified framework for 3D pose forecasting from 2D images or historical motions. In this work, we integrate MLPs into 3D multi-person motion prediction for the first time, and propose novel lightweight yet effective local-global and spatial-temporal interaction operations for enhanced representation learning.

3. Method

Given the historical 3D observation of multi-person joint sequences $X = \{x_t^p\}_{1:T}^{1:P} \in \mathbb{R}^{3J \times P \times T}$, which consists of P persons, T time frames, and J joints per person, our goal is to predict the future joint sequences $Y = \{y_t^p\}_{1:T'}^{1:P}$ for the next T' frames. In this paper, we propose the Efficient Multi-Person Motion Prediction (EMPMP) network

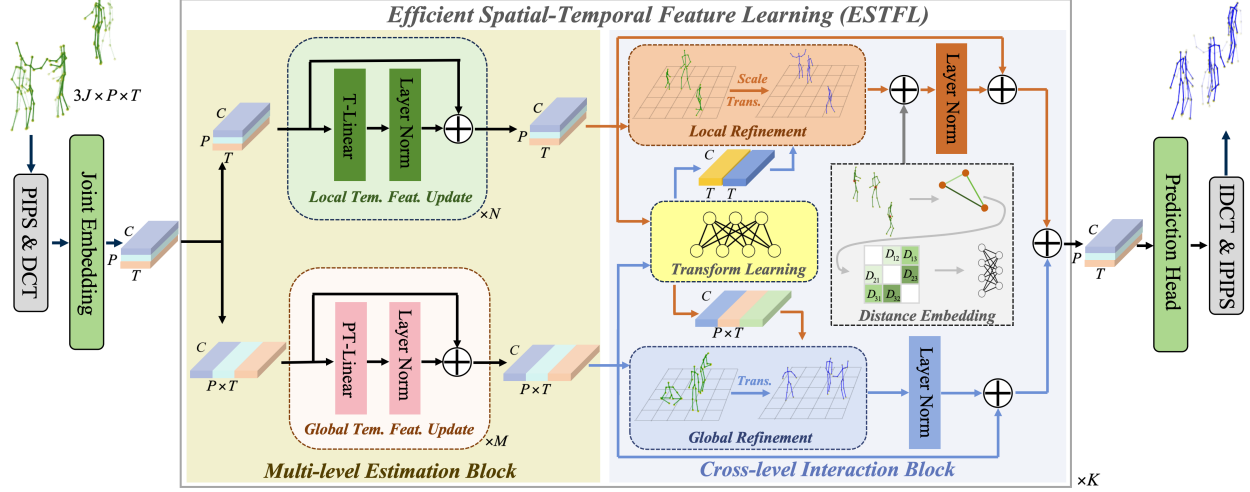


Figure 2. Framework of our EMPMP network. The input is mapped to order-agnostic feature space by PIPS and DCT stages. The Joint Embedding and ESTFL stages are designed to capture spatial and temporal representations as well as their interactions via lightweight architectures. The Prediction Head, IDCT and IPIPS stages are used to map the representation to predicted motion.

with lightweight yet expressive spatial and temporal interactions for 3D multi-person motion prediction. The framework is presented in Sects. 3.1- 3.4, with network training and implementation details in Sect. 3.5.

3.1. Overall Framework

Our EMPMP network conducts 3D multi-person motion prediction through a cascaded pipeline as shown in Fig. 2. The framework comprises three core modules: the Joint Embedding for encoding skeletal geometry, the Efficient Spatial-Temporal Feature Learning to model multi-level spatial and temporal interactions, and the Prediction Head for future motion prediction mapping, as detailed in the subsequent subsections.

To resolve order ambiguity of individuals, we introduce Permutation-Invariant Person Sorting (PIPS), which reorders individuals by descending pairwise distances in the first frame, yielding order-agnostic features $X' \in \mathbb{R}^{3J \times P \times T}$ (please refer to supplementary material for details). Following this, we apply Discrete Cosine Transform (DCT) to compress temporal dynamics into a low-frequency spectrum $X'' = DX'$, where $D \in \mathbb{R}^{T \times T}$ is the DCT coefficient matrix conducted on temporal dimension.

3.2. Joint Embedding

The joint embedding stage transforms the DCT-processed input X'' into a high-dimensional feature space through a linear projection, integrating spatial joint information into a structured representation:

$$Z = X''W_0 + b_0, \quad Z \in \mathbb{R}^{C \times P \times T}, \quad (1)$$

where $W_0 \in \mathbb{R}^{3J \times C}$ and $b_0 \in \mathbb{R}^C$ are learnable parameters shared across all individuals (P) and frames (T). This

embedding integrates spatial features into a shared feature space, enabling holistic pose modeling rather than isolated joint processing.

3.3. Efficient Spatial-Temporal Feature Learning

The Efficient Spatial-Temporal Feature Learning (ESTFL) stage forms the core of our framework, designed to capture inter-person spatial dependencies and inter-frame temporal dynamics within the sequential multi-person representation $Z \in \mathbb{R}^{C \times P \times T}$. It comprises two main components: the *Multi-level Estimation* (ME) block and *Cross-level Interaction* (CI) block, described as follows.

3.3.1. Multi-level Estimation (ME) Block

The ME block is proposed to capture temporal information for both individual and multi-person representations. A dual-stream architecture is employed to capture both local (individual) and global (scene-level) dynamics.

Local Temporal Feature Update. The local stream focuses on capturing individual motion patterns through iterative temporal feature updates. In each iteration, a shared temporal linear transformation (T-linear) is applied across all persons, followed by layer normalization and residual connection. Formally, the i -th iteration of local update is:

$$Z_l^i = Z_l^{i-1} + \beta(Z_l^{i-1}W_l^i + b_l^i), \quad i = 1, \dots, N, \quad (2)$$

where $Z_l^0 = Z$, $W_l^i \in \mathbb{R}^{T \times T}$ and $b_l^i \in \mathbb{R}^T$ are learnable parameters, and β denotes layer normalization. By sharing these parameters across all persons, this iterative process progressively refines fine-grained individual motion understanding while preserving computational efficiency.

Global Temporal Feature Update. The global stream is designed to learn temporal interactions across all individ-

uals within the scene. We merge the person (P) and temporal (T) dimensions of Z to form a global representation $Z_g^0 \in \mathbb{R}^{C \times PT}$, which encodes dynamics of the entire scene. A linear transformation along the person-temporal dimension (PT-linear) is applied with a residual connection, and the global features are iteratively refined by:

$$Z_g^j = Z_g^{j-1} + \gamma(Z_g^{j-1}W_g^j + b_g^j), \quad j = 1, \dots, M, \quad (3)$$

where $W_g^j \in \mathbb{R}^{PT \times PT}$ and $b_g^j \in \mathbb{R}^{PT}$ are learnable parameters for the j -th iteration, and γ is layer normalization. By progressively refining the global representation, the global stream effectively captures the temporal dependencies across all individuals in the scene.

The multi-level features of $Z_l^N \in \mathbb{R}^{C \times P \times T}$ and $Z_g^M \in \mathbb{R}^{C \times PT}$ are subsequently integrated in the Cross-level Interaction (CI) block, enabling a comprehensive interaction between individual motion details and global scene dynamics for enhanced motion prediction.

3.3.2. Cross-level Interaction (CI) Block

The CI block facilitates efficient information exchange between local and global representations. For simplicity, we omit the superscripts and denote the local and global features as Z_l and Z_g , respectively.

Local Refinement with Global Representation. The global representation Z_g captures high-level temporal interactions across all individuals in the scene, providing essential contextual information to enhance local features of each individual. To refine local representations, we apply affine transformations with scale $S \in \mathbb{R}^{C \times T}$ and translation $H \in \mathbb{R}^{C \times T}$ learned from Z_g via linear transformations:

$$S = Z_gW_S + b_S, \quad H = Z_gW_H + b_H, \quad (4)$$

where $W_S, W_H \in \mathbb{R}^{PT \times T}$ and $b_S, b_H \in \mathbb{R}^T$ are learnable parameters, and they are employed to integrate global context across individuals and temporal frames. To maintain consistent global influence across all individuals, S and H are shared across persons, yielding expanded representations $\hat{S}, \hat{H} \in \mathbb{R}^{C \times P \times T}$. The refined local representation is then computed by applying these transformations:

$$Z_l' = \rho(Z_l \odot (1 + \hat{S}) + \hat{H}), \quad (5)$$

where $\odot$ denotes the Hadamard (element-wise) product, and ρ is layer normalization. This process dynamically adjusts local features using global context, ensuring that the refined representations capture both individual motion details and scene-level temporal dynamics.

Inter-person Distance Embedding. To further refine the local representation, we incorporate the spatial relationships between individuals in the scene by computing an inter-person distance matrix $D \in \mathbb{R}^{P \times P \times T}$, which captures the

spatial interactions between individuals at each time frame. The matrix element is defined as

$$D_{p_1 p_2 t} = \|x_{t, \text{hip}}^{p_1} - x_{t, \text{hip}}^{p_2}\|_2, \quad (6)$$

where $D_{p_1 p_2 t}$ represents the Euclidean distance between the hips of the p_1 -th and p_2 -th persons at time t . This distance matrix is then linearly transformed along its first dimension to produce the distance embedding:

$$\tau(D) = DW_D + b_D, \quad (7)$$

where $W_D \in \mathbb{R}^{P \times C}$ and $b_D \in \mathbb{R}^C$ are learnable parameters. The embedded distance $\tau(D) \in \mathbb{R}^{C \times P \times T}$ is then used to refine the local representation:

$$Z_l^* = Z_l + \xi(Z_l' + \tau(D)), \quad (8)$$

where ξ represents layer normalization. This process ensures that the local representation incorporates both individual motion dynamics and the spatial relationships between individuals, thereby improving the model's accuracy in predicting multi-person motions.

Global Refinement with Local Representation. To refine the global representation using local information from $Z_l \in \mathbb{R}^{C \times P \times T}$, we first compress the local features across individuals and time into a compact representation for affine transformation. Specifically, we combine the P and T dimensions of Z_l , resulting in $Z_{l2g} \in \mathbb{R}^{C \times PT}$, and apply a linear transformation to learn the global translation:

$$G = Z_{l2g}W_G + b_G, \quad (9)$$

where $W_G \in \mathbb{R}^{PT \times PT}$, $b_G \in \mathbb{R}^{PT}$ are learnable parameters. Based on above learned translation, the refined global representation is computed as

$$Z_g^* = Z_g + \pi(Z_g + G), \quad (10)$$

where π is layer normalization. This process ensures that the global representation is enriched with detailed local dynamics while maintaining stable feature distributions.

Finally, the refined local and global representations are combined into a unified feature:

$$Z^* = Z_l^* + \alpha \hat{Z}_g^*, \quad Z^* \in \mathbb{R}^{C \times P \times T}, \quad (11)$$

where $\hat{\cdot}$ denotes dimension transform for compatibility, and α is the combination parameter.

The CI block enables bidirectional information exchange between local and global representations, significantly improving the model's ability to capture complex motion dynamics while ensuring computational efficiency.

The ESTFL stage employs a dual-block architecture, consisting of the ME block for multi-level feature learning and the CI block for cross-level information exchange. This

straightforward yet effective design allows the model to capture multi-level spatial and temporal interactions, boosting its ability to predict motion dynamics. Experimental results demonstrate the superior performance of above designs, showing substantial improvements in both accuracy and efficiency for 3D multi-person motion prediction task.

3.4. Prediction Head and Loss Function

Given $Z^* \in \mathbb{R}^{C \times P \times T'}$, the prediction head is designed to map the high-dimensional features to joint-wise representations through linear transformation in two dimensions:

$$\tilde{Z} = Z^* W_T + b_T, \quad \hat{Z} = \tilde{Z} W_C + b_C, \quad (12)$$

where $W_T \in \mathbb{R}^{T \times T'}$, $b_T \in \mathbb{R}^{T'}$ and $W_C \in \mathbb{R}^{C \times 3J}$, $b_C \in \mathbb{R}^{3J}$ are learnable parameters that perform the transformations in temporal (T) and feature (C) dimensions, respectively. The final joint sequence is obtained by applying Inverse DCT cross the temporal dimension as $Y' = D^{-1} \hat{Z}$, followed by Inverse PIPS to restore the order and generate a permutation-invariant prediction $Y \in \mathbb{R}^{3J \times P \times T'}$.

The loss function consists of two components including the mean joint loss and the velocity loss as utilized in [17]:

$$\mathcal{L} = \mathcal{L}^{Joint} + \mathcal{L}^{Vel}. \quad (13)$$

The mean joint loss quantifies the discrepancy between the predicted and ground-truth joint sequences by:

$$\mathcal{L}^{Joint} = \frac{1}{P \cdot T' \cdot J} \sum_{p=1}^P \sum_{t=1}^{T'} \sum_{j=1}^J \|y_t^{p,j} - \hat{y}_t^{p,j}\|^2, \quad (14)$$

where $y_t^{p,j} \in \mathbb{R}^3$ represents the predicted world coordinates of the j -th joint for the p -th person at the t -th frame, and $\hat{y}_t^{p,j}$ is the corresponding ground truth. The velocity loss accounts for the predicted and ground-truth velocities between consecutive frames:

$$\mathcal{L}^{Vel} = \frac{1}{P \cdot (T' - 1) \cdot J} \sum_{p=1}^P \sum_{t=1}^{T'-1} \sum_{j=1}^J \|v_t^{p,j} - \hat{v}_t^{p,j}\|^2, \quad (15)$$

where the predicted and ground-truth velocities are computed as follows:

$$v_t^{p,j} = y_{t+1}^{p,j} - y_t^{p,j}, \quad \hat{v}_t^{p,j} = \hat{y}_{t+1}^{p,j} - \hat{y}_t^{p,j}. \quad (16)$$

By incorporating both joint and velocity constraints, the loss function enforces consistency in both spatial and temporal dynamics, thereby enhancing the accuracy and realism of the motion prediction.

3.5. Network Training and Implementation Details

Our network is built on the framework proposed in Sects. 3.1- 3.4. For the hyper-parameters, we set the combination parameter $\alpha = 0.2$ in Eq. (11), and we take

$K = 4$ stages of ESTFL to efficiently capture the spatial and temporal features. Each ESTFL stage consists of $N = 16$, $M = 1$ iterations for local and global temporal feature updates, respectively. The feature channel dimension is set to $C = 45$ for experiments on CMU-Mocap and MuPoTs-3D datasets, and $C = 39$ for experiments on the 3DPW dataset.

The network is trained with batch size 128 and learning rate 3×10^{-4} by Adam optimizer. During training, we apply data augmentation after PIPS stage, with random vertical-axis rotations and person order permutations as [41]. All experiments are conducted on a single NVIDIA 3090 GPU. Our code will be released if the paper is accepted.

4. Experiments

In this section, we first present the dataset settings and evaluation metrics, followed by comparisons of the experimental results. Additionally, we conduct ablation studies to investigate the effectiveness of our designs.

4.1. Datasets and Evaluation Metrics

3DPW and AMASS. The 3DPW dataset [42] is a large-scale collection containing over 51,000 frames of human activities in unconstrained environments, with scenes involving two individuals. The AMASS dataset [29] is a comprehensive motion capture dataset, featuring over 40 hours of motion data from 300+ subjects performing more than 11,000 distinct motions. These datasets are used in four distinct settings: 1) *3DPW-Ori*: As in [41, 47], the input consists of 1030ms (16 frames) and the prediction horizon is 900ms (14 frames) based on the original 3DPW dataset. 2) *3DPW-RC*: This setting is similar to *3DPW-Ori*, but with camera displacements removed to enhance the realism of human motion, as in [47]. 3) *AMASS/3DPW-Ori*: The model is pre-trained on AMASS, then fine-tuned and evaluated on *3DPW-Ori*, with pre-training and fine-tuning configurations matching those in [41, 47]. 4) *AMASS/3DPW-RC*: Similar to *AMASS/3DPW-Ori*, but the model is fine-tuned and evaluated on *3DPW-RC* instead, as in [47].

CMU-Mocap. The CMU-Mocap dataset [1] consists of 140 subjects and 2,605 sequences featuring one or two individuals. Following [45], 6,000 synthetic scenes with three individuals are generated, each consisting of 60 frames at a frame rate of 15 FPS. The *CMU-Syn* setting includes three motion prediction tasks: predicting 1 second of motion from 1-second input (1s/1s) for short-term prediction, predicting 2 seconds from 2-second input (2s/2s) for long-term prediction, and predicting 3 seconds from 1-second input (1s/3s) for extended forecasting. For the 1s/1s setting, we randomly sample 30 frames from 60 during training, increasing the training samples to enhance model robustness.

MuPoTS-3D. The MuPoTS-3D dataset [32] contains over 8,000 frames across 20 real-world scenes with up to three

Metric	Settings	3DPW-Ori	3DPW-RC	CMU-Syn		AMASS/ 3DPW-Ori	AMASS/ 3DPW-RC	CMU-Syn/MuPoTS	
	In/Out Length	1030ms / 900ms		2s / 2s	1s / 1s	1030ms / 900ms		2s / 2s	1s / 1s
MPJPE	MRT [45] ²⁰²¹	167.1	134.2	160.7	83.5	145.3	124.8	181.0	93.1
	SoMoFormer [41] ²⁰²²	150.0	107.7	148.3	77.0	114.6	96.4	162.0	89.1
	TBIFormer [33] ²⁰²³	149.9	124.1	169.7	91.0	140.4	114.3	173.4	93.2
	JRT [47] ²⁰²³	154.9	114.0	157.6	84.3	131.9	102.3	161.5	93.0
	T2P [20] ²⁰²⁴	137.5	110.1	136.4	75.9	121.1	100.4	163.6	94.7
	Ours	131.8	99.5	128.0	73.5	119.9	92.1	160.1	92.4
VIM	MRT [45] ²⁰²¹	66.9	55.2	61.5	36.3	59.2	52.3	70.1	41.6
	SoMoFormer [41] ²⁰²²	57.8	44.4	56.7	34.4	46.3	40.0	63.7	39.3
	TBIFormer [33] ²⁰²³	60.2	51.2	64.6	40.3	56.4	47.4	68.3	44.0
	JRT [47] ²⁰²³	58.5	47.0	56.6	34.0	47.2	39.5	64.7	41.2
	T2P [20] ²⁰²⁴	56.3	46.9	54.6	34.2	49.4	42.4	64.0	42.0
	Ours	53.3	41.3	50.2	32.4	48.6	38.4	62.9	40.6
APE	MRT [45] ²⁰²¹	125.1	123.6	95.5	60.0	115.3	110.8	135.1	81.2
	SoMoFormer [41] ²⁰²²	118.3	114.9	87.1	55.2	97.5	101.4	130.8	80.6
	TBIFormer [33] ²⁰²³	115.9	115.5	96.9	63.6	110.3	108.0	137.7	84.4
	JRT [47] ²⁰²³	123.0	120.5	95.8	61.6	114.1	112.5	125.7	78.4
	T2P [20] ²⁰²⁴	115.1	115.1	94.5	59.9	110.1	108.7	147.7	92.8
	Ours	98.6	96.6	83.2	52.3	95.0	90.6	124.9	76.2

Table 1. Performance comparisons across multiple settings using MPJPE, VIM, and APE metrics. Our method achieves superior performance with significantly fewer parameters. The best result is highlighted in bold.

subjects. Similar to strategy in [45], for the *CMU-Syn/MuPoTS* setting, the models are trained on *CMU-Syn*, and performance is evaluated by predicting 15 frames (1s) and 30 frames (2s) of motion sequences on MuPoTS-3D dataset.

Evaluation Metrics. We evaluate model performance using metrics of Mean Per Joint Position Error (MPJPE) [19], Visibility-Ignored Metric (VIM) [2] and Aligned Mean Per Joint Position Error (APE). MPJPE measures the mean Euclidean distance between predicted and ground truth 3D joint positions across all frames. VIM assesses the per-frame accuracy by computing mean joint-coordinate distance. APE removes global displacement, focusing on local posture prediction. For these metrics, lower values indicate better performance. Due to space constraints, we report average values for key frames in APE and VIM. Further details on selection scheme and additional metrics are available in the supplementary material.

4.2. Experimental Results

We evaluate our approach on 3D multi-person motion prediction tasks in three scenarios: standard, incorporating pre-training, and cross-dataset, to comprehensively assess the model’s effectiveness and generalizability.

Results on 3DPW-Ori and 3DPW-RC. As shown in the first and second columns of Tab. 1, our model outperforms previous methods in both *3DPW-Ori* and *3DPW-RC* settings for standard multi-person motion prediction tasks, demonstrating significant superiority. The substantial reductions in MPJPE, VIM, and APE highlight the effec-

tiveness of our approach. Notably, our model achieves state-of-the-art performance with only 0.04M parameters, as shown in Tab. 2, which constitutes less than approximately 1.2% of the parameter count of prior models. Additionally, our model requires less GPU memory and computational FLOPs, and more detailed comparisons on computational costs can be found in the supplementary material.

Settings	3DPW	3DPW- Pretrain	CMU-Syn & MuPoTS	
In/Out Length	1030ms / 900ms		2s / 2s	1s / 1s
MRT [45]	6.55M	6.55M	6.29M	6.29M
SoMoFormer [41]	12.91M	12.91M	12.91M	12.91M
TBIFormer [33]	3.39M	3.39M	7.31M	4.24M
JRT [47]	3.68M	3.68M	3.68M	3.68M
T2P [20]	4.50M	4.50M	4.60M	4.50M
Ours	0.04M	0.65M	0.17M	0.05M

Table 2. Model Parameters (in Millions) for different settings. 3DPW refers to the settings of *3DPW-Ori* and *3DPW-RC*, 3DPW-Pretrain covers *AMASS/3DPW-Ori* and *AMASS/3DPW-RC*, while CMU-Syn&MuPoTS include *CMU-Syn* and *CMU-Syn/MuPoTS*.

Results on CMU-Syn. As demonstrated in the third and fourth columns of Tab. 1, our model outperforms the compared methods across all three evaluation metrics for both short-term (1s/1s) and long-term (2s/2s) motion prediction in the *CMU-Syn* setting, while maintaining a lightweight design with less parameters as shown in Tab. 2. Furthermore, the results for the extended forecasting task (1s/3s) are presented in Tab. 3, where our EMPMP consistently sur-

Metric	MPIPE			VIM			APE		
Out Length	1s	2s	3s	1s	2s	3s	1s	2s	3s
MRT [45]	89	145	189	65	97	128	86	102	111
SoMoFormer [41]	84	151	202	66	108	140	81	101	112
TBIFormer [33]	98	165	218	72	114	148	88	103	116
JRT [47]	88	154	206	60	105	147	89	106	118
T2P [20]	89	139	177	65	89	117	87	103	116
Ours	84	130	167	60	85	113	77	93	103

Table 3. Performance comparison in *Mocap-Syn* (1s/3s) setting.

passes the other methods within the 3s prediction horizon, particularly for long-range predictions at 2s and 3s.

Results on AMASS/3DPW-Ori and AMASS/3DPW-RC with Pre-training. For tasks involving a pre-training strategy, we utilize a 0.65M-parameter model to enhance our modeling capability, ensuring a fair comparison (please refer to the supplementary material for further details). As shown in Tab. 1, our model achieves state-of-the-art performance in four out of six comparisons, while remaining highly competitive in the other two.

Results on CMU-Syn/MuPoTS for Cross-Dataset Evaluation. As shown in the last two columns of Tab. 1, our model excels in cross-dataset evaluation, achieving the best performance in the 2s/2s setting while using fewer parameters. For the 1s/1s setting, our model attains the best APE and competitive MPJPE/VIM results, demonstrating strong generalization capabilities.

4.3. Ablation Study

In this subsection, we conduct detailed ablation studies to assess the effectiveness of our designs and the impact of hyper-parameters. All of the models are evaluated by MPJPE in the *3DPW-Ori* and *CMU-Syn* settings.

Effectiveness of Multi-level Estimation. The ME block integrates individual and multi-person features via local and global temporal feature update mechanisms. As shown in the first and second rows of Tab. 4, removing either the local temporal feature update (LTFU) or the global temporal feature update (GTFU) stream results in a performance decline compared to the network that includes both components, as seen in the third row. These comparisons underscore the importance of both streams and highlight the effectiveness of the proposed ME block.

Effectiveness of Cross-level Interaction. The CI block enhances local and global representations through interaction via learned affine transformations (LRwGR and GRwLR). As shown in Tab. 4, removing the entire CI block (third row), or individually removing either the local or global refinements (fourth and fifth rows), significantly reduces performance compared to the model with CI block (last two rows). This clearly validates the critical role of each component. Notably, we learn both scale and translation parameters during local refinement, whereas global refinement em-

plies only translation. Further ablation studies in the supplementary material confirm the effectiveness of these design choices in managing affine transformations.

LTFU	GTFU	LRwGR	GRwLR	DE	MPJPE
×	✓	×	×	×	160.4 (0.04M)
✓	×	×	×	×	143.4 (0.07M)
✓	✓	×	×	×	140.9 (0.10M)
✓	✓	×	✓	×	137.7 (0.12M)
✓	✓	✓	×	×	133.8 (0.14M)
✓	✓	✓	✓	×	129.8 (0.16M)
✓	✓	✓	✓	✓	128.0 (0.17M)

Table 4. Ablation study on different components in the *CMU-Syn* (2s/2s) setting. **LTFU**: Local Temporal Feature Update; **GTFU**: Global Temporal Feature Update; **LRwGR**: Local Refinement with Global Representation; **GRwLR**: Global Refinement with Local Representation; **DE**: Inter-person Distance Embedding. The number of parameters (in Millions) is shown in parentheses.

Effectiveness of Distance Embedding. In the CI block, we explicitly integrate inter-person distance embedding (DE) to refine local representations by capturing spatial relationships between individuals. As illustrated in the last two rows of Tab. 4, the model without the distance embedding operation exhibits decreased performance compared to our full model, validating the effectiveness of this component.

Learn Multi-level feat.	Learn Cross-level feat.	MPJPE	Params	FLOPs
Self-Attention	CI block	131.7	3.28M	39.83G
ME block	Cross-Attention	133.6	0.26M	3.07G
ME block	CI block	128.0	0.17M	1.67G

Table 5. Comparisons with Attention-based design in *CMU-Syn* (2s/2s) setting. Network with our designs achieves superior performance with fewer parameters and lower computational cost.

Comparison with Attention-based Design. In our model, multi-level temporal features are learned through linear layers and layer normalization within the ME block, whereas cross-level interactions are achieved through learned affine transformations in the CI block. By contrast, previous models have predominantly employed Attention-based mechanisms for both feature learning and interaction. Besides comparing our method against traditional Attention-based architectures in Sect. 4.2, we further investigate the effectiveness of our ME/CI block by replacing them with Attention-based designs (details in supplementary material), as shown in Tab. 5. Results indicate that Attention-based design incurs higher computational costs and yields inferior MPJPE scores. Conversely, our proposed design achieves superior performance while substantially reducing the number of parameters and computational FLOPs.

Impact of Hyper-parameters. We examine the impact of essential hyper-parameters, including the number of local/global temporal feature update iterations (N/M) and the

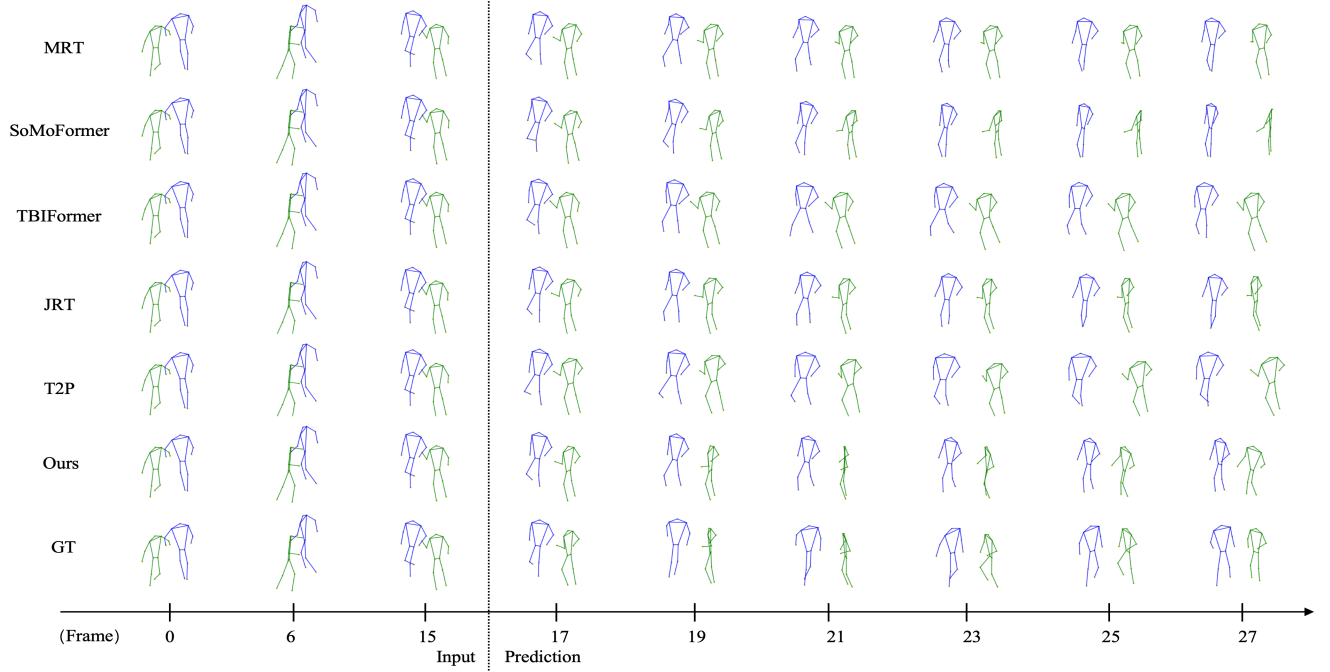


Figure 3. Qualitative results in the *3DPW-RC* setting involve predicting 14 frames using 16 frames of input. Different colors are used to distinguish between individuals in the sequence.

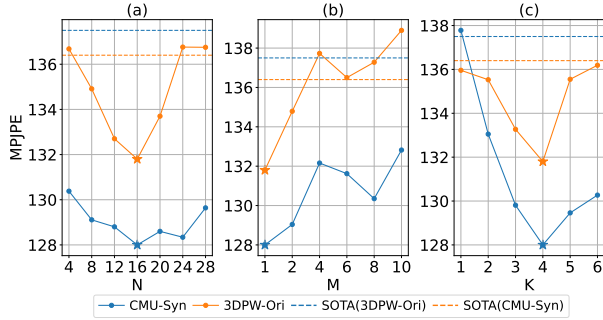


Figure 4. Performance curves with varied hyper-parameters in *CMU-Syn* (2s/2s) and *3DPW-Ori* (1030ms/900ms) settings. N/M represent the iteration number of local/global temporal feature updates per ME block, and K denotes the number of ESTFL stages.

number of ESTFL stages (K). As illustrated in Fig. 4, the optimal configuration ($N = 16, M = 1, K = 4$) achieves the best performance for both the *CMU-Syn* (2s/2s) and *3DPW-Ori* (1030ms/900ms) settings. Our EMPMP model outperforms state-of-the-art methods across most hyper-parameter variations, highlighting its robustness for hyper-parameters. Notably, our network with $M \leq 2$ presents consistently superior performance than cases of $M > 2$, indicating that excessive global iterations may lead to over-smoothing of temporal features and thus degrade prediction accuracy. Further ablation studies with different values of combination parameter α in Eq. (11) are detailed in supplementary material, with $\alpha = 0.2$ producing the best results.

4.4. Qualitative Results

In this section, we provide a qualitative comparison between our model and existing methods using the *3DPW-RC* setting, as illustrated in Fig. 3. The ground truth consists of two subjects performing distinct actions: the left subject remains largely static with minimal leg movement, while the right subject executes a full turn, exhibiting more dynamic motion. Our model outperforms previous methods in predicting the movements of both subjects. For the left subject, our model preserves subtle details of the pose with greater precision. For the right subject, our model captures the dynamic motion while ensuring stability, maintaining temporal coherence and realism throughout the movement. Additional visualization results demonstrating the effectiveness of our model are provided in supplementary material, further highlighting the superiority of our EMPMP network.

5. Conclusion

In this paper, we introduce EMPMP, an efficient architecture for 3D multi-person motion prediction. Our design uses a dual-branch structure to capture multi-level spatial-temporal features and a cross-level interaction strategy for information exchange, significantly reducing parameters and computational complexity. Experiments across multiple datasets validate our approach’s efficiency and accuracy. Future work will explore stochastic motion prediction and advanced feature fusion strategies.

Acknowledgment This work was supported by NSFC with grant numbers of 62306167, 12125104, 12426313. It is also funded by project ZR2024QA161 and ZR2023QG143 supported by Shandong Provincial Natural Science Foundation.

References

- [1] Cmu graphics lab motion capture database. <http://mocap.cs.cmu.edu/>. 5
- [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. 6
- [3] Nasir Ahmed, T Natarajan, and Kamisetty Ramamohan Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1):90–93, 1974. 2
- [4] Elmira Amirloo, Amir Rasouli, Peter Lakner, Mohsen Rohani, and Jun Luo. Latentformer: Multi-agent transformer-based interaction modeling and trajectory prediction. *arXiv preprint arXiv:2203.01880*, 2022. 1
- [5] Jose Daniel Azofeifa, Julieta Noguez, Sergio Ruiz, José Martín Molina-Espinosa, Alejandra J Magana, and Bedrich Benes. Systematic review of multimodal human-computer interaction. *Informatics*, 9(1):13, 2022. 1
- [6] Aritz Badiola-Bengoa and Amaia Mendez-Zorrilla. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. *Sensors*, 21(18):5996, 2021. 1
- [7] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. *arXiv preprint arXiv:2207.00499*, 2022. 2
- [8] Clementine Calba, Flavie L Goutard, Linda Hoinville, Pascal Hendriks, Ann Lindberg, Claude Saegeman, and Marisa Peyre. Surveillance systems evaluation: a systematic review of the existing approaches. *BMC Public Health*, 15:1–13, 2015. 1
- [9] Shoufa Chen, Enze Xie, Chongjian Ge, Runjian Chen, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021. 2
- [10] Xiaobo Chen, Huanjia Zhang, Yu Hu, Jun Liang, and Hai Wang. Vnagt: Variational non-autoregressive graph transformer network for multi-agent trajectory prediction. *IEEE Transactions on Vehicular Technology*, 72(10):12540–12552, 2023. 1
- [11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 2
- [12] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1423–1432, 2019. 2
- [13] Andrew Clark, Anban W Pillay, and Deshendran Moodley. A system for pose analysis and selection in virtual reality environments. In *Conference of the South African Institute of Computer Scientists and Information Technologists*, pages 210–216, 2020. 1
- [14] 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*, pages 11467–11476, 2021. 2
- [15] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4346–4354, 2015. 1, 2
- [16] Qiang Fu, Xingui Zhang, Jinxiu Xu, and Haimin Zhang. Capture of 3d human motion pose in virtual reality based on video recognition. *Complexity*, 2020(1):8857748, 2020. 1
- [17] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023. 2, 5
- [18] Bassem HadjKacem, Walid Ayedi, Mohamed Abid, and Hichem Snoussi. Multi-shot human re-identification for the security in video surveillance systems. In *IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pages 203–208, 2018. 1
- [19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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. 6
- [20] Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1617–1628, 2024. 1, 2, 6, 7
- [21] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *Proceedings of the European conference on computer vision*, pages 89–107, 2025. 1
- [22] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014. 1, 2
- [23] Wenhao Li, Mengyuan Liu, Hong Liu, Tianyu Guo, Ti Wang, Hao Tang, and Nicu Sebe. Graphmlp: A graph mlp-like architecture for 3d human pose estimation. *Pattern Recognition*, 158:110925, 2025. 2
- [24] Yungche Li, Chingtang Chang, Chinchang Cheng, and Yulen Huang. Baseball swing pose estimation using openpose. In *IEEE International Conference on Robotics, Automation and Artificial Intelligence*, pages 6–9, 2021. 1
- [25] Hai Liu, Tingting Liu, Zhaoli Zhang, Arun Kumar Sangaiah, Bing Yang, and Youfu Li. Arhpe: Asymmetric relation-aware representation learning for head pose estimation in in-

- dustrial human–computer interaction. *IEEE Transactions on Industrial Informatics*, 18(10):7107–7117, 2022. 1
- [26] Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, Yaonan Wang, and Rainer Stiefelhagen. Transkd: Transformer knowledge distillation for efficient semantic segmentation. *arXiv preprint arXiv:2202.13393*, 2022. 2
- [27] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019. 1, 2
- [28] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 1, 2
- [29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 5
- [30] 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. 1, 2
- [31] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European conference on computer vision*, pages 474–489, 2020. 1, 2
- [32] 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 *IEEE International Conference on 3D Vision*, pages 120–130, 2018. 5
- [33] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17121–17130, 2023. 1, 2, 6, 7
- [34] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43:1–54, 2015. 1
- [35] Xiangshi Ren, Chaklam Silpasuwanchai, and John Cahill. Human-engaged computing: the future of human–computer interaction. *CCF Transactions on Pervasive Computing and Interaction*, 1:47–68, 2019. 1
- [36] Hyocheol Ro, Yoon Jung Park, Jung-Hyun Byun, and Tack-Don Han. Display methods of projection augmented reality based on deep learning pose estimation. In *ACM SIGGRAPH Posters*, pages 1–2. 2019. 1
- [37] Tyler Rose, Chang S Nam, and Karen B Chen. Immersion of virtual reality for rehabilitation-review. *Applied Ergonomics*, 69:153–161, 2018. 1
- [38] Kyuhong Shim, Iksoo Choi, Wonyong Sung, and Jungwook Choi. Layer-wise pruning of transformer attention heads for efficient language modeling. In *IEEE 18th International SoC Design Conference*, pages 357–358, 2021. 2
- [39] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 1
- [40] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 2
- [41] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezaatofghi. Somoformer: Multi-person pose forecasting with transformers. *arXiv preprint arXiv:2208.14023*, 2022. 1, 2, 5, 6, 7
- [42] 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*, pages 601–617, 2018. 5
- [43] Hongsong Wang, Jian Dong, Bin Cheng, and Jiashi Feng. Pvrnet: A position-velocity recurrent encoder-decoder for human motion prediction. *IEEE Transactions on Image Processing*, 30:6096–6106, 2021. 2
- [44] Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. *Advances in Neural Information Processing Systems*, 18, 2005. 2
- [45] 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, 5, 6, 7
- [46] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023. 1
- [47] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. Joint-relation transformer for multi-person motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9816–9826, 2023. 1, 2, 5, 6, 7
- [48] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022. 2
- [49] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *CVPR*, 2022. 1