scientific reports

Check for updates

OPEN

Origin centric and part based pose decomposition for 3D human pose estimation

Zhijie Lin[1,7]✉, Jinxin Yao[1,7]✉, Juan Huang[2], Jingjing Chen[3,7], Yingying Xu[4]✉, Lu Yang[5], Lei Zhao[6] & Wei Xing[6]

Transformer-based approaches have recently made significant advancements in 3D human pose estimation from 2D inputs. Existing methods typically either consider the entire 2D skeleton for global features extraction or break it into independent parts for local features learning. However, capturing the spatial dependencies of the entire 2D skeleton does not effectively facilitate learning local spatial features, while partitioning the skeleton into independent segments disrupts the relevance of individual joints to the whole. In this paper, we propose a novel Origin-centric Part Transformer (OPFormer) block to address this issue through two steps: Skeleton Separation and Skeleton Recombination. Skeleton Separation separates the 2D skeleton into several distinct parts, enabling the extraction of fine-grained local spatial features that accurately reflect the geometric structure of the human body. Secondly, we introduce the concept of a human skeleton Origin, which serves as a central hub to reconnect different parts through Skeleton Recombination. The resulting local features, when fused with global features from the Spatial Transformer Encoder, yield more accurate 3D results. Comprehensive experiments conducted on the Human3.6M and MPI-INF-3DHP benchmark datasets verify that our approach attains state-of-the-art performance. It should be emphasized that OPFormer achieves a Mean Per Joint Position Error (MPJPE) of 37.6mm on the Human3.6M dataset without any additional training data.

Human pose estimation (HPE) is a critical and challenging task in the field of computer vision (CV), with significant applications in areas such as human-robot interaction[1], virtual reality[2], and action recognition[3]. The goal of HPE is to predict the position of each joint from input images or videos. Depending on whether the predicted joint contains depth information, HPE can be categorized into 2D human pose estimation (2D HPE) and 3D human pose estimation (3D HPE). With the advancement of deep learning technology, the field of 2D HPE has matured considerably. The accuracy and generalization of 2D detectors[4–9] have reached an advanced level. However, the output of these models is limited to 2-dimensional information. In contrast, although 3D HPE faces more challenges, the addition of depth information enables it to provide richer 3D spatial information and a better understanding of human movements and interactions. Consequently, the 2D-to-3D methods of applying the developed 2D detector to the 3D HPE tasks hold significant potential as a monocular solution.

The inherent ambiguity of monocular data allows a single 2D pose to correspond to multiple possible 3D poses, making it challenging to accurately recover 3D poses from single-frame 2D joint position information. Recently, driven by Transformer[10] for its ability to capture long-distance dependencies, 2D-to-3D methods[11–18] leveraging video frame sequences with temporal motion information have made significant progress. Starting with an input video, the 2D-to-3D methods first detect the 2D keypoints and subsequently infer the 3D joint positions based on the detected 2D keypoints. Among these, PoseFormer[11] captures global spatial dependencies from the entire 2D human skeleton and models temporal features from frame sequences to output accurate 3D poses. MixSTE[13] further separates the entire human skeleton into multiple joints to model more fine-grained temporal features. However, previous methods overlook the fact that not all human joints are closely related in spatial position, thus capturing the spatial dependencies of the entire 2D skeleton does not effectively facilitate the learning of spatial features.

[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China. [2]School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China. [3]Faculty of Science, Hong Kong Baptist University, Hong Kong 999077, China. [4]School of Humanities and Social Science, Beihang University, Beijing 100083, China. [5]Office of Education Affair, China Jiliang University, Hangzhou 310018, China. [6]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. [7]Zhijie Lin, Jinxin Yao and Jingjing Chen contributed equally to this work. ✉email: linzhijie@zust.edu.cn; 222203855010@zust.edu.cn; yingxu21@buaa.edu.cn

To address the above issue, we performed a MixSTE performance analysis in modeling spatial relationships among 17 joints from the Human3.6M[19] dataset. We begin by averaging the attention weights of the Spatial Self-Attention modules in MixSTE, resulting in a 17×17 attention map. Based on the distribution patterns of high-weight regions in the average attention map, the 2D skeleton can be segmented into five joint groups and an Origin, as illustrated in Fig. 1a. Each joint group is represented by a distinct color for clarity, and the corresponding average attention weights between the joints in these groups are visualized with matching colors in Fig. 1b. In addition to the joint groups shown in Fig. 1a, b also illustrates the average attention weights between the hip joint and all other joints (denoted as "Origin"), as well as the average attention weights among all joints outside of the defined joint groups (denoted as "Other"). As shown by the average attention weights associated with the "Other" label in Fig. 1b, joints outside the defined groups show low spatial dependencies, due to their lack of direct anatomical connections and their tendency to move independently. In contrast, joints within the same part exhibit high spatial dependencies, as reflected by the "Right Leg" value, due to their close geometric connections and their synchronized movement. Furthermore, the results indicate that all joints are strongly associated with the hip joint, the root of the human body, and its position subsequently affects the spatial positions of all other joints.

Building on the aforementioned observations, achieving accurate and fine-grained spatial feature modeling is essential for improving the accuracy of 3D human pose estimation. In this paper, we introduce a new Origin-centric and part-based pose decomposition method for more precise spatial feature modeling. Specifically, we propose a novel Origin-centric Part Transformer (OPFormer) block to model fine-grained dependencies through two steps: Skeleton Separation and Skeleton Recombination.

First, the introduction of Skeleton Separation aims to focus the model's attention on more localized spatial features. This approach is motivated by our observation, as illustrated in Fig. 1a, that spatial dependencies between closely connected joints exhibit higher correlation compared to more distant joints. In light of this, we hypothesize that by isolating these finer-grained local relationships, we can better capture the specific movements and interactions within each body segment. To this end, we divide the 2D skeleton into distinct, independent parts through the process of Skeleton Separation. Each of these parts, such as the legs or arms, is treated as a separate unit, allowing the model to capture local dependencies within each part individually and more effectively. This division not only simplifies the spatial modeling but also improves the granularity of the features learned from closely related joints.

Secondly, we introduce the concept of the hip joint as the Origin of the human skeleton. This Origin plays a critical role as a spatial hub that maintains the overall coherence of the skeleton by reconnecting the separated parts. As shown in Fig. 1b, through the process of Skeleton Recombination, each part is recombined with the Origin, forming what we refer to as Origin-centric Parts (OParts). This process ensures that all parts are not treated as independent isolated regions, but as components whose spatial relationships are maintained relative to the entire body. By establishing this hierarchical relationship, where each OPart remains anchored to the hip joint, we preserve the structural integrity of the skeleton while modeling local spatial dependencies.

Furthermore, although some joint pairs do not belong to the same part, they still exhibit a certain degree of dependency. We adopted the strategy of previous works[11–14,18] to model relationships outside of parts and generate global spatial features that complement local spatial features. Consequently, our proposed OPFormer employs a parallel structural design, with two parallel channels responsible for capturing global and local spatial features, respectively. Finally, after fusing the outputs of the two channels, the temporal block is utilized to obtain the temporal features of each joint over multiple time steps, thereby generating more accurate 3D pose results.

The main contributions of this paper can be summarized in three aspects:

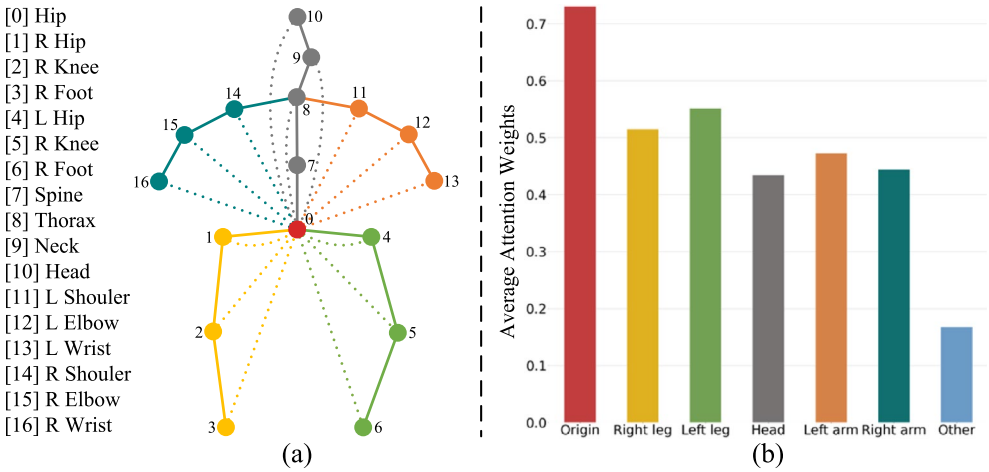

Fig. 1. (**a**) The origin-centric part diagram. The 2D skeleton is separated into multiple parts and an Origin based on the distribution of spatial dependencies, with the corresponding colors from (**b**) used to draw the joints and connections. (**b**) The average dependencies within different joint groups in (**a**). The "Origin" value indicates the average dependency of all joints on the hip joint, while the "Other" value indicates the average dependency between all joints outside the joint groups defined in (**a**).

- We proposed a novel Origin-centric Part Transformer (OPFormer), which models fine-grained local dependencies within each body part through part-level decomposition and utilizes the Origin joint as a central hub to preserve global structural coherence across all parts of the skeleton.
- We developed an alternative network structure with a dual-channel parallel mechanism to capture spatial features across various ranges. It is complemented by a temporal block to capture temporal features of different joints, and finally it enhanced 3D pose estimation accuracy.
- We evaluated the proposed model with two benchmark datasets: Human3.6M and MPI INF-3DHP, and the proposed model reached SOAT results without using any extra training data.

Related work

Monocular 3D human pose estimation

Monocular 3D human pose estimation (3D HPE) primarily consists of end-to-end methods and 2D-to-3D methods. Earlier works focused on directly inferring 3D poses from single-view images. Li et al.[20] first used convolutional neural networks (CNN[21]) to regress 3D pose and explored the ability of CNN to encode the relationships of human structure. However, end-to-end methods do not leverage the superior performance of 2D detectors. Chen et al.[22] introduced the 2D-to-3D method, dividing 3D HPE into two stages: first estimating a 2D pose and then estimating its depth by matching to a library of 3D poses. However, predicting the 3D pose from the 2D pose cannot avoid inherent depth ambiguity, necessitating additional information for accurate 3D pose estimation. Zhang et al.[23] used commercial LiDAR to explore the 3D background and scan neighbors, providing spatial and contextual cues for individual point clouds to improve 3D HPE performance. But hardware-based methods are costly and have stringent environmental requirements, making them challenging to apply in real-world scenarios. Other methods rely on video sequences and use multi-frame image data to predict 3D pose. Earlier video-based methods employed recurrent neural networks (RNN[24]) to process time series, whereas VideoPose3D[25] used a dilated temporal convolution network to process sequences connected by 2D joint coordinates. Compared with RNN-based methods, CNN-based methods can process time series data in parallel. Similar to CNNs, Transformers[10] can also process sequence data in parallel and effectively capture long-distance dependencies. Recently, Transformers have become fundamental in natural language processing (NLP), and their application[26,27] to visual tasks has achieved advanced performance.

Transformer-based methods

Due to the adoption of self-attention mechanism, Transformer can capture global dependencies across 2D pose sequence. StridedFormer[14] reduces the redundancy of 2D pose sequences by replacing the full connection layer in the Vanilla Transformer Encoder (VTE) with Strided Convolution, thereby reducing the data dimension layer by layer. PoseFormerV2[17] expands the receptive field using frequency domain representation from 2D pose sequences. MotionBERT[18] learns human motion representations from large and heterogeneous datasets for downstream tasks like 3D HPE after fine-tuning. MotionAGFormer[16] uses Transformer to capture global information and graph convolutional networks (GCNs) to capture local information, ensuring a balanced and comprehensive representation of human motion. In addition, recent methods[28,29] leverage Transformer as the backbone of the denoiser within diffusion architectures to achieve more accurate 3D pose estimation. The above methods are rough to use the entire 2D skeleton as the source of spatial features of the model, but we observe that not all human joints are closely related during movement.

Part-aware methods

Some work has also attempted to separate the entire skeleton to investigate new schemes to improve the performance of models. HEMlets Pose[30] uses heatmaps of the three joints to represent the relative depth information of the end joints of each part to shorten the gap between 2D observation and 3D interpretation. SRNet[31] divides human body into local areas to solve the problem of long-tailed distribution caused by a small number of pose samples. Anatomy3D[32] decomposes 3D pose estimation into bone direction and length prediction, leveraging skeletal consistency and a fully convolutional architecture to enhance temporal modeling without recurrent units. PoseAug[33] introduced part-aware Kinematic Chain Space to evaluate the rationality of local joint angle for enhanced pose. HSTFormer[34] captures temporal features hierarchically, treating joints, parts, and the entire human body as distinct objects to analyze from local to global scales. STCFormer[35] divides human body into static parts and dynamic parts, and generates two different Structure-enhanced Positional Embeddings to add to the network structure. While these methods separate the human body into several independent parts, they often neglect the relationships between the parts and the whole. In contrast, our proposed method introduces an Origin-centric Part representation that not only captures fine-grained local dependencies within each part but also preserves global spatial coherence by anchoring all parts to a shared reference joint. This design enables more comprehensive spatial modeling and leads to improved pose estimation accuracy.

Methods

The objective of our network is to lift a 2D skeleton sequence generated by a 2D detector into a 3D pose sequence. Figure 2 illustrates the overall network architecture, which consists of five key components. The Origin-centric Part Transformer (OPFormer) Block and the Spatial Transformer Encoder are combined in a dual-channel parallel structure to model spatial dependencies at multiple scales within a single-frame 2D pose. The OPFormer Block, introduced in this work, focuses on capturing intra-part relationships among joints and generating localized spatial features. In line with previous studies[11–14,18], the Spatial Transformer Encoder is employed to model inter-joint relationships across the entire body, producing global spatial features. The outputs of these two channels are adaptively fused to strengthen the network's ability to capture spatial dependencies

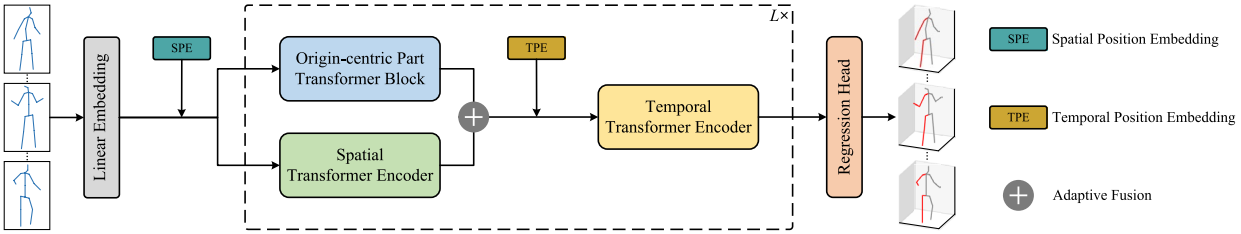

Fig. 2. Overview of the network architecture. The proposed network is designed with L stacked loops. Within each loop, the Origin-centric Part Transformer Block and the Spatial Transformer Encoder are arranged in a parallel structure to facilitate feature fusion. Subsequently, these components are integrated with the Temporal Transformer Encoder in a serial structure to enable feature transmission.

at different scales. The fused spatial features are subsequently passed to the Temporal Transformer Encoder, which models temporal dependencies of each joint across sequential frames. The spatial and temporal modeling processes are performed alternately, and the 3D pose sequence is obtained through a Regression Head after L loops. We start with an overview of the computational workflow and then provide a comprehensive explanation of each network component.

Network architecture

As shown in Fig. 2, the input of the model is a 2D skeleton sequence $X \in \mathbb{R}^{T\times J\times C_i}$ with a confidence score, where T is the number of frames, J is the number of joints and C_i is the number of input channels. The Linear Embedding layer is initially used to project the input to a high-dimensional feature $F^0 \in \mathbb{R}^{T\times J\times C_e}$, and then the learnable spatial position embedding $P_S \in \mathbb{R}^{1\times J\times C_e}$ is added. After inputting the prepared features into the parallel channel, we utilize the OPFormer Block to compute the local spatial features $F^l_{PS} \in \mathbb{R}^{T\times J\times C_e}$ $(l = 1, \ldots, L)$, and utilize the Spatial Transformer Encoder to compute the global spatial feature $F^l_{GS} \in \mathbb{R}^{T\times J\times C_e}$ $(l = 1, \ldots, L)$, where L is the depth of the network and C_e is the number of channels with embedded features. We use adaptive fusion to fuse the output of the two channels to generate a complete spatial feature $F^l_S \in \mathbb{R}^{T\times J\times C_e}$ $(l = 1, \ldots, L)$, this process is defined as:

$$F^l_S = \alpha^l_{PS} \odot F^l_{PS} + \alpha^l_{GS} \odot F^l_{GS}, \tag{1}$$

where $\odot$ is element-wise production, the adaptive fusion weights α^l_{PS} and α^l_{GS} is defined as:

$$\alpha^l_{PS}, \alpha^l_{GS} = Softmax(Concat(F^l_{PS}, F^l_{GS})W_A), \tag{2}$$

where W_A is a learnable linear transformation.

Then, the temporal position embedding $P_T \in \mathbb{R}^{T\times 1\times C_e}$ is added to the fused feature F^l_S and input into the Temporal Transformer Encoder to compute the temporal feature $F^l_T \in \mathbb{R}^{T\times J\times C_e}$ $(l = 1, \ldots, L)$. Finally, a linear Regression Head is applied to F^L_T to estimate the final 3D pose $Y \in \mathbb{R}^{T\times J\times 3}$, which contains the coordinate values of the three dimensions in the spatial coordinate system.

Origin-centric part transformer block

The OPFormer Block is designed to compute the local spatial features of a single-frame 2D pose. Unlike previous methods[11,13,14] that primarily focus on the global spatial correlations of all human joints, our approach introduces a more localized scope for capturing spatial dependencies.

Skeleton separation and skeleton recombination

As illustrated in Fig. 1a, our method separates the human skeleton into several parts based on spatial dependencies: right leg, left leg, trunk, right arm, and left arm, along with an Origin. Specifically, as shown in Fig. 3, the complete skeleton feature $F_S \in \mathbb{R}^{J\times C_e}$ is segmented into five part features $F^p_P \in \mathbb{R}^{J_p\times C_e}$ $(p = 1, \ldots, 5)$ and an Origin feature $F_O \in \mathbb{R}^{1\times C_e}$ via Skeleton Separation, where J_p denotes the number of joints in the p-th part. Due to stronger spatial dependencies within each part, joints belonging to the same part can provide more fine-grained local spatial representations. However, they lack the implicit positional relationships between different parts. To address this, we recombine each part with the Origin joint to form multiple Origin-centric Parts (OParts). More precisely, for each part, we concatenate the Origin feature $F_O \in \mathbb{R}^{1\times C_e}$ with $F^p_P \in \mathbb{R}^{J_p\times C_e}$ through Skeleton Recombination to generate the OPart feature $F^p_{OP} \in \mathbb{R}^{(J_p+1)\times C_e}$. The process of skeleton separation and recombination can be denoted as:

$$F_O, F^1_P, \ldots, F^5_P = Split(F_S), \tag{3}$$

$$F^p_{OP} = Concat(F_O, F^p_P), \tag{4}$$

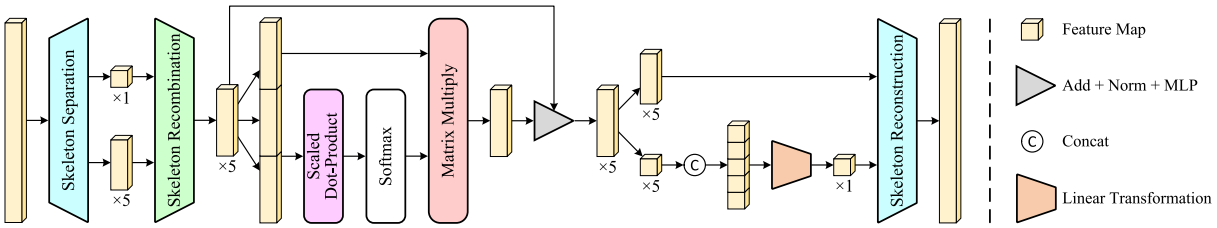

Fig. 3. Architecture and computational flow of the origin-centric part transformer block. The input 2D skeleton feature undergoes two key processes: skeleton separation and skeleton recombination, resulting in the generation of OPart features. After each OPart feature is projected into a higher-dimensional space, spatial dependencies are modeled using scaled dot-product attention. Since the origin features span multiple OParts and carry positional dependencies with each part, they are converged into a single feature through a linear transformation module. Finally, the complete skeleton feature is restored via skeleton reconstruction.

Local spatial dependencies modeling

With Origin and joints of the same part included, each OPart is treated as a unit to capture local spatial dependencies, and F_{OP}^{p} is then fed into the Scaled Dot-Product Attention (*SDPA*) mechanism. For computing the query matrix Q_{OP}, the key matrix K_{OP} and the value matrix V_{OP}, we project F_{OP}^{p} by Linear layer:

$$Q_{OP} = F_{OP}^{p} W_{OP}^{Q}, K_{OP} = F_{OP}^{p} W_{OP}^{K}, V_{OP} = F_{OP}^{p} W_{OP}^{V}, \tag{5}$$

where W_{OP}^{Q}, W_{OP}^{K}, and W_{OP}^{V} are projection matrices. The query matrix Q_{OP}, the key matrix K_{OP} and the value matrix V_{OP} are then fed into the Scaled Dot-Product Attention (*SDPA*):

$$SDPA(Q_{OP}, K_{OP}, V_{OP}) = Softmax\left(\frac{Q_{OP}(K_{OP})^{T}}{\sqrt{d_K}}\right) V_{OP}, \tag{6}$$

where T is the matrix transpose operation, d_K is the number of dimensions of the key K_{OP}.

To model relations inside each Origin-centric Part, multi-head self-attention (*MSA*) mechanism is applied:

$$\begin{aligned} &MSA(Q_{OP}, K_{OP}, V_{OP}) = Concat\left(head_1, \ldots, head_H\right) W_{OP}^{O}, \\ &\text{where} \quad \text{head}_h = SDPA\left(Q_{OP}^{h}, K_{OP}^{h}, V_{OP}^{h}\right), \quad h = 1, \ldots, H, \end{aligned} \tag{7}$$

where W_{OP}^{O} is a learnable linear transformation, H is the number of attention heads. The complete process of local spatial dependencies modeling can be defined as follows:

$$F_{OP}^{p}{}' = MSA(LN(F_{OP}^{p})) + F_{OP}^{p}, \tag{8}$$

$$Z_{OP}^{p} = MLP(LN(F_{OP}^{p}{}')) + F_{OP}^{p}{}', \tag{9}$$

where *MSA* represents the multi-head attention, *MLP* represents the multilayer perceptron and *LN* is the layer normalization layer.

Skeleton reconstruction

$$F_{PS} = Concat(Concat\left(Z_{O}^{1}, \ldots, Z_{O}^{5}\right) W_{O}, Z_{P}^{1}, \ldots, Z_{P}^{5}), \tag{10}$$

where W_O is a learnable linear transformation that converges multiple Origin features into a single feature. Through these steps, both local and global spatial dependencies can be accurately captured and used to generate the final 3D pose estimation.

Transformer encoder

Spatial transformer encoder

We utilize Spatial Transformer Encoder (STE) to capture the spatial relationships between the joints outside the parts of a single frame 2D skeleton. The Spatial Multi-Head Self-Attention (*SMSA*) treats individual joints from the entire body as tokens. Each attention head is computed in a way that follows the scaled dot-product attention. The definition of *SMSA* and scaled dot-product attention is as follows:

$$SMSA\left(Q_S, K_S, V_S\right) = Concat\left(head_1, \ldots, head_H\right) W_S^{O}, \tag{11}$$

$$\text{head}_h = Softmax\left(\frac{Q_S^{h}\left(K_S^{h}\right)^{T}}{\sqrt{d_K}}\right) V_S^{h}, \tag{12}$$

where W_S^O is a learnable linear transformation, H is the number of attention heads, T is the matrix transpose operation, d_K is the number of dimensions of the key K_S^h. For computing the query matrix Q_S, the key matrix K_S and the value matrix V_S, we project spatial feature F_S by Linear layer:

$$Q_S^h = F_S W_S^{(Q,h)}, K_S^h = F_S W_S^{(K,h)}, V_S^h = F_S W_S^{(V,h)}, \tag{13}$$

where $W_S^{(Q,h)}$, $W_S^{(K,h)}$, and $W_S^{(V,h)}$ are projection matrices of the *h*-th head.

Temporal transformer encoder

We utilize a Temporal Transformer Encoder (TTE) to capture the temporal relationships between different time steps of a single joint. The Temporal Multi-Head Self-Attention (*TMSA*) treats individual joints from different frames as tokens. Each attention head is computed in a way that follows the scaled dot-product attention. The definition of *TMSA* and scaled dot-product attention is as follows:

$$TMSA\left(Q_T, K_T, V_T\right) = Concat\left(head_1, \ldots, head_H\right) W_T^O, \tag{14}$$

$$\text{head}_h = Softmax\left(\frac{Q_T^h \left(K_T^h\right)^T}{\sqrt{d_K}}\right) V_T^h, \tag{15}$$

the definition of each parameter in the formula is similar to that in the Spatial Transformer Encoder.

Regression head

A linear Regression Head is employed to estimate the final 3D pose $\hat{Y} \in \mathbb{R}^{T\times J\times 3}$. We then compute the loss between $\hat{Y}$ and GT to train the network from end to end. Since the proposed model processes 2D skeleton sequences at joint-level and frame-level, the final loss function $\mathscr{L}$ is composed of two elements:

$$\mathscr{L} = \sum_{t=1}^{T}\sum_{j=1}^{J}\left\|\hat{Y}_{t,j} - Y_{t,j}\right\|_2 + \lambda_M \sum_{t=2}^{T}\sum_{j=1}^{J}\left\|\hat{M}_{t,j} - M_{t,j}\right\|_2, \tag{16}$$

where $\hat{M}_t = \hat{Y}_t - \hat{Y}_{t-1}$, $M_t = Y_t - Y_{t-1}$, and the constant coefficient λ_M is used to balance position accuracy and motion smoothness.

Experiments

We evaluated our proposed OPFormer model with two benchmark datasets: Human3.6M[19] and MPI-INF-3DHP[37].

Dataset and evaluation metrics

Human3.6M is a commonly utilized indoor dataset for 3D human pose estimation, comprising 3.6 million video frames. It features 11 professional actors engaged in 15 diverse everyday activities, including actions such as eating, sitting, and walking. Each subject was filmed from four different angles. We followed the protocol of previous works[11,25,32,38], training the model on subjects 1, 5, 6, 7, and 8, and testing on subjects 9 and 11.

MPI-INF-3DHP is a comprehensive dataset for 3D human pose estimation, featuring both indoor and outdoor environments. It consists of over 1.3 million frames, capturing 8 categories of activities performed by 8 subjects from 14 camera angles, including a greater variety of poses.

For the Human3.6M dataset, we use MPJPE (Mean Per Joint Position Error) and P-MPJPE (Procrustes-aligned Mean Per Joint Position Error) as metrics to evaluate the performance of OPFormer. MPJPE calculates the average Euclidean distance between the estimated joint and the ground truth. P-MPJPE, on the other hand, computes the MPJPE following rigid alignment between the estimated joint positions and the ground truth, providing greater robustness to single joint prediction errors. We adopt these proven metrics (Percentage of Correct Keypoint with 150 mm, and Area Under the Curve) to evaluate the proposed model in the MPI-INF-3DHP dataset.

Implementation details

Our approach is implemented with PyTorch[39] and trained and tested on an NVIDIA RTX 4080 GPU. To ensure a fair comparison, we select specific 2D input lengths for different datasets: Human3.6M (T=27, 81, 243) and MPI-INF-3DHP (T=81). For Human3.6M, we use both the 2D predictions of the Stacked Hourglass[7] and the 2D ground truth as inputs to the proposed model, following[16,18]. Following[11,13], the 2D ground truth data is utilized for the MPI-INF-3DHP dataset. During model training, we set the input channel C_i=3 (including the coordinates of x and y along with confidence scores) following[18,40] and use random horizontal flipping as data augmentation following[25,32,38]. The model is trained for 120 epochs using the Adam[41] with an initial learning rate of 0.0002 and a weight decay of 0.99. To address the issue of limited video memory, we use a gradient accumulation strategy. Specifically, we accumulate gradients over batches of size 4 and update the model weights after every 8 such accumulations.

	T	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MPJPE																	
VideoPose3D[25]	243	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
PoseFormer[11]	81	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Stridedformer[14]	351	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
MixSTE[13]	243	36.7	39.0	36.5	39.4	**40.2**	**44.9**	39.8	36.9	**47.9**	54.8	39.6	37.8	39.3	29.7	30.6	39.8
STCFormer[35]	243	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	*26.2*	27.7	40.5
MotionBERT[18]	243	36.3	38.7	38.6	33.6	42.1	50.1	*36.2*	35.7	50.1	56.6	41.3	*37.4*	37.7	**25.6**	*26.5*	39.2
HDFormer[36]	96	**34.7**	41.7	*36.0*	38.4	41.1	*45.3*	39.6	37.4	49.0	63.1	*39.8*	38.9	40.2	29.3	29.1	40.3
MotionAGFormer[16]	243	36.8	*38.5*	**35.9**	*33.0*	41.1	48.6	38.0	*34.8*	49.0	**51.4**	40.3	*37.4*	*36.3*	27.2	27.2	*38.4*
Ours	243	*35.3*	**37.4**	37.2	**32.5**	*40.4*	48.1	**35.8**	**34.4**	*48.4*	*52.6*	**39.0**	**36.2**	**35.7**	**25.6**	**25.6**	**37.6**
P-MPJPE																	
VideoPose3D[25]	243	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
PoseFormer[11]	81	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Stridedformer[14]	351	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	35.1	25.0	35.2
MixSTE[13]	243	30.8	33.1	*30.3*	31.8	33.1	39.1	31.1	30.5	42.5	**44.5**	34.0	30.8	32.7	*22.1*	22.9	32.6
STCFormer[35]	243	*29.3*	33.0	30.7	30.6	*32.7*	38.2	29.7	**28.8**	42.2	*45.0*	*33.3*	**29.4**	31.5	**20.9**	**22.3**	*31.8*
MotionBERT[18]	243	30.8	32.8	32.4	28.7	34.3	38.9	30.1	30.0	42.5	49.7	36.0	30.8	*22.0*	31.7	23.0	32.9
HDFormer[36]	96	**27.9**	32.8	**29.7**	30.6	**32.5**	**35.0**	**28.9**	*29.2*	**38.3**	50.0	**32.9**	30.1	31.8	23.6	*22.8*	**31.7**
MotionAGFormer[16]	243	31.0	*32.6*	31.0	**27.9**	34.0	38.7	31.5	30.0	41.4	45.4	34.8	30.8	31.3	22.8	23.2	32.5
Ours	243	29.9	**31.6**	31.6	*28.0*	33.5	*38.1*	*29.4*	29.5	*40.5*	46.8	34.0	*29.8*	**21.6**	30.6	**22.3**	*31.8*

Table 1. Quantitative comparison results under MPJPE (mm) and P-MPJPE (mm) with state-of-the-art methods on Human3.6M using the detected 2D pose as input. *T* denotes the number of input frames. The top two results are marked in bold and italics, respectively.

MPJPE	*T*	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
VideoPose3D[25]	243	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.2
PoseFormer[11]	81	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Stridedformer[14]	351	27.1	29.4	26.5	27.1	28.6	33.0	30.7	26.8	38.2	34.7	29.1	29.8	26.8	19.1	19.8	28.5
MixSTE[13]	243	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
HDFormer[36]	96	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	21.6
STCFormer[35]	243	20.8	21.8	20.0	20.6	23.4	25.0	23.6	*19.3*	27.8	26.1	21.6	20.6	19.5	14.3	15.1	21.3
MotionAGFormer[16]	243	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	17.3
MotionBERT[18]	243	*16.7*	*19.9*	*17.1*	*16.5*	*17.4*	*18.8*	*19.3*	20.5	*24.0*	*22.1*	*18.6*	*16.8*	*16.7*	**10.8**	*11.5*	*17.8*
Ours	243	**14.2**	**15.4**	**15.0**	**14.0**	**15.5**	**17.7**	**16.3**	**16.2**	**21.0**	**20.0**	**16.2**	**13.7**	**8.9**	*14.2*	**9.7**	**15.2**

Table 2. Quantitative comparison results using MPJPE (mm) with state-of-the-art methods on Human3.6M, using 2D ground truth key points as input.

Comparison with state-of-the-art

Table 1 shows the results of our comparison with other methods, including the error and average error for all 15 actions. For the sake of a clear and intuitive presentation of the results, we select only the best results of models without considering variants. To ensure a fair comparison, only the results of models trained without additional pre-training on external data are considered. With the same detector of Stacked Hourglass, our method achieved the best result of average MPJPE of 37.6mm, which outperforms MotionAGFormer[16] by 0.8mm (2.4%) MPJPE and MotionBERT[18] by 1.6mm (4.8%) MPJPE. It is worth mentioning that our average error is only 0.1mm worse than the pre-trained MotionBERT with extra data. Moreover, OPFormer achieved the best results to date in 9 of the 15 categories. A P-MPJPE of 31.8mm was obtained, which outperforms MotionAGFormer by 0.7mm (2.1%) MPJPE and only 0.1mm worse than HDFormer.

To further evaluate the performance of OPFormer, we used 2D ground truth keypoints as input, allowing for a direct comparison with state-of-the-art methods. The results presented in Table 2 show that by eliminating the error introduced by the 2D pose estimation process, our method achieves an average MPJPE of 15.2mm. This result significantly outperforms all other methods and marks a substantial improvement of 2.6mm (14.6%) over MotionBERT. This considerable enhancement highlights the effectiveness of OPFormer in accurately estimating 3D poses when given precise 2D input data. The removal of 2D estimation errors underscores the inherent robustness and precision of our approach. The experiments conducted with both 2D detected keypoints and 2D ground truth keypoints collectively demonstrate the versatility and reliability of our method across different

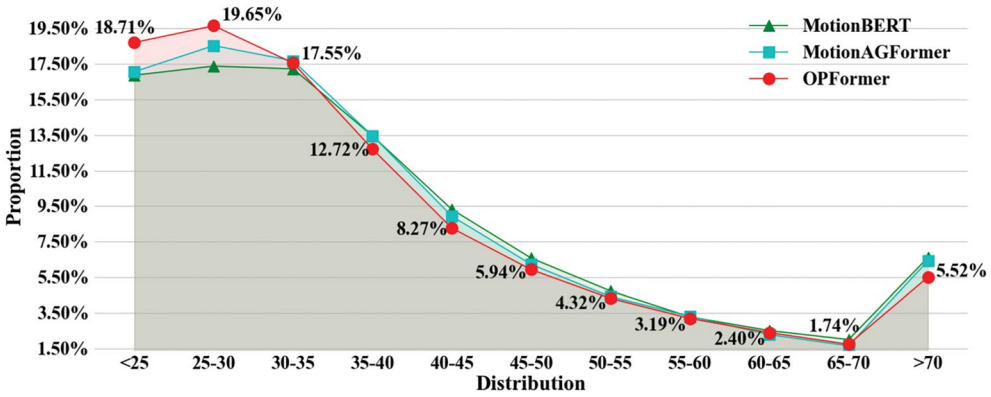

Fig. 4. The MPJPE distribution on the testset S9 and S11 in Human3.6M. The x-axis and y-axis represents the error interval and poses proportion in a certain interval.

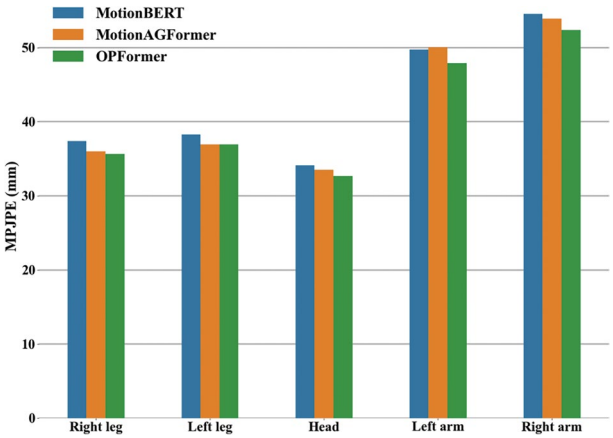

Fig. 5. The average joint error for each part across all frames in the Human3.6M testset.

input types. These findings solidify OPFormer as a highly competitive model in the field of 3D human pose estimation.

Furthermore, Fig. 4 shows the MPJPE distribution comparisons between OPFormer and two other recent methods, including MotionBERT and MotionAGFormer. With the same input of 2D keypoints from Stacked Hourglass, OPFormer leads to the highest poses proportion with MPJPE less than 30mm, and the lowest poses proportion with MPJPE larger than 40mm. The results demonstrate that our method has fewer errors for difficult actions and higher accuracy for simple actions.

In Fig. 5, we also compare the average joint error for each part with MotionBERT and MotionAGFormer. As shown, the results indicate that the left arm and the right arm exhibit higher detection errors due to more complex movements. The proposed method demonstrates superior accuracy, as evidenced by its performance on each part, particularly in the left and right arms, highlighting its effectiveness in precisely locating joints and handling complex movements.

To verify the generalization of OPFormer on other datasets, we compared it with other methods on MPI-INF-3DHP dataset. As shown in Table 3, with 2D ground truth keypoints as input and a frame number of 81, the proposed model achieves the best results in three evaluation metrics with PCK of 98.7%, AUC of 85.3% and MPJPE of 15.9mm. On MPJPE in particular, the OPFormer is 0.3mm (1.8%) lower than the best previous method MotionAGFormer[16]. On the other two metrics, our approach also matched the to-data best reported results. These findings demonstrate that OPFormer possesses superior generalization ability and an enhanced capacity for spatial feature modeling when applied to more complex datasets. This robustness across different datasets underscores the effectiveness of OPFormer in 3D human pose estimation.

Ablation study

In order to analyze the impact and performance of each component of the model in depth, we conducted a series of ablation experiments to evaluate their effectiveness. The Stacked Hourglass detector provides 2D keypoints based on the Human3.6M dataset.

As shown in Table 4, since our 3D HPE model is based on video sequences, the Temporal Encoder is a fundamental component. For a fair comparison, we incrementally add modules to the baseline model, which

Method	*T*	PCK↑	AUC↑	MPJPE↓
VideoPose3D[25]	81	86.0	51.9	84.0
PoseFormer[11]	9	88.6	56.4	77.1
MixSTE[13]	27	94.4	66.5	54.9
HDFormer[36]	96	**98.7**	72.9	37.2
STCFormer[35]	81	**98.7**	83.9	23.1
MotionAGFormer[16]	81	98.2	**85.3**	16.2
Ours	81	**98.7**	**85.3**	**15.9**

Table 3. Quantitative comparison results on MPI-INF-3DHP using the 2D ground truth as input.

	Temporal Encoder	Spatial Encoder	Part Block	Origin-centric design	MPJPE (mm)
Baseline	✓		✓		40.4
	✓	✓			38.4
	✓	✓	✓		38.3
Ours	✓	✓	✓	✓	37.6

Table 4. Ablation study for each component used in our method on Human3.6M with MPJPE (mm).

Depth	Dimension	Frames	MPJPE (mm)
6	512	243	38.0
8	128	243	40.5
8	256	243	39.6
8	512	27	43.9
8	512	81	40.4
8	512	243	37.6
10	512	243	37.9

Table 5. Ablation study on different hyper-parameters in our method on Human3.6M with MPJPE (mm).

Origin	MPJPE (mm)
Thorax	38.6
Spine	38.1
Hip	37.6

Table 6. Ablation study of different Origin choices in our method on Human3.6M with MPJPE (mm).

initially includes only the Temporal Encoder, and calculate the MPJPE losses to assess the effect of each component. When only the Spatial Encoder or a simple Part Block is added to the model, the results indicate that more accurate 3D pose estimations cannot be achieved using solely global or local spatial features. However, when we combine these three components, the addition of the Part Block alone only reduces the MPJPE by 0.1 mm, demonstrating that independent local spatial features bring little improvement. The best results (37.6 mm MPJPE) are achieved when we introduce the Origin-centric design to form the OPFormer Block. This addition reduces MPJPE by 2.8mm (6.9%) compared to the baseline model, validating the efficacy of our design.

Table 5 shows how the setting of different hyperparameters impacts the model performance under MPJPE. The proposed model has three main hyperparameters: the depth of the network stack, the dimension of the model, and the number of input frames. To assess the effect of each configuration on the model, we vary each hyper-parameter across three different values while keeping the other two fixed. Referring to the results in Table 5, we select the configuration with Depth=8, Dimension=512, and Frames=243. Table 6 presents the results of an ablation study evaluating the impact of different joint selections as the Origin in our Origin-centric framework. This can be attributed to the hip joint's anatomical role as the root of the human skeleton, providing a more stable and centrally located reference for modeling spatial dependencies.

Qualitative results

Attention visualization: we visualize the average spatial attention maps for the different layers of the OPFormer Block (top) and the STE (bottom), as illustrated in Fig. 6. From the bottom, the results indicate that the dependencies between all human joints and the hip joint are greatly reduced compared to MixSTE[13]. As expected, the Spatial Transformer Encoder (SPE) focuses on the spatial relationships of joints outside the part, whereas the dependencies inside the part and the implicit position relationships based on the Origin are transferred to the OPFormer Block. From the top, it is evident that layers 1 to 5 of the OPFormer Block mainly capture the dependencies between different joints within the part, whereas layers 6 to 8 primarily capture the dependencies between the internal joints and the Origin. This clear division of modeling different types of dependencies is due to the OPFormer block and the overall network architecture design. By segmenting the human body structure based on the Origin, our double-channel parallel structure is capable of capturing more fine-grained global and local spatial features. This enhances the model's ability to comprehensively understand and model the spatial dependencies inherent in human poses.

Qualitative analysis: Figure 7 presents a qualitative comparison between OPFormer and two recent methods: MotionBERT[18] and MotionAGFormer[16]. The examples are selected from the Photo and SittingDown actions in the Human3.6M test set. As highlighted in the blue circled areas, the poses estimated by OPFormer exhibit a closer resemblance to the ground truth compared to the other methods. Importantly, for complex actions such as "SittingDown," our method achieves significantly more accurate results. This indicates that OPFormer maintains competitive performance even in challenging action categories, demonstrating its robustness and precision in 3D human pose estimation.

Conclusion

We propose the OPFormer model for spatial feature modeling that leverages Skeleton Separation and Skeleton Recombination for effective monocular 3D pose estimation. The OPFormer uniquely captures localized joint dependencies within distinct body parts while preserving the structural relevance of each joint to the entire body. Integrated into an innovative network architecture, the OPFormer operates within a parallel and alternating framework, efficiently capturing and fusing spatio-temporal features to enhance performance. Comprehensive experiments on benchmark datasets validate that the proposed model substantially improves the accuracy of 3D HPE, demonstrating robust performance across varying motion patterns. However, our proposed method

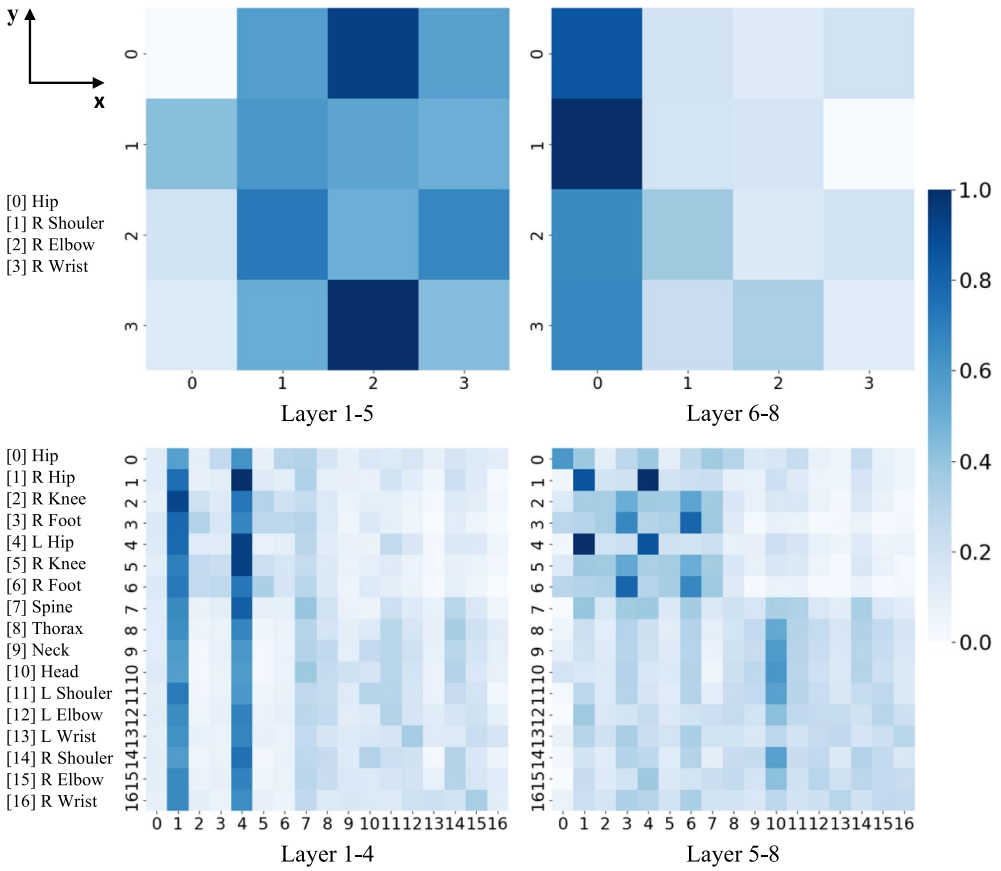

Fig. 6. Visualization of attention maps in the OPFormer block (top) and the STE (bottom). The x-axis and y-axis correspond to the joint indexes in the part and body. The label indicates which layers the average attention comes from.

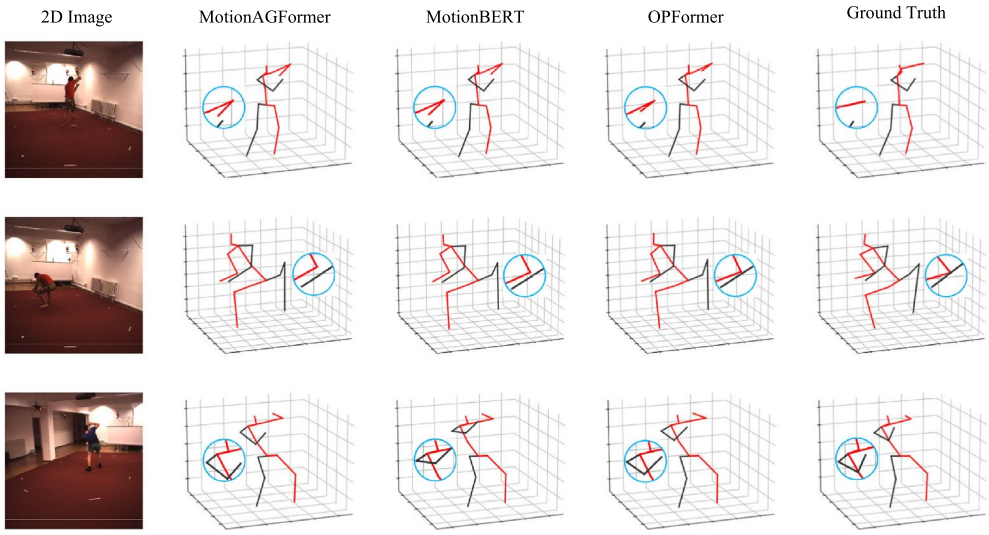

Fig. 7. Qualitative comparison between OPFormer and the state-of-the-art methods on Human3.6M testset. Red represents the torso and the left, and black indicates the right of the estimated body. The blue circle indicates areas where our method outperforms others.

incurs a relatively higher computational cost compared to some existing approaches. In future work, we plan to explore efficiency-aware variants of our model to reduce computational cost while preserving high accuracy, potentially by sharing Origin embeddings across parts or integrating more efficient attention mechanisms.

Data availability

The data used in this study is sourced from publicly available datasets. Human3.6M dataset URL: http://vision.imar.ro/human3.6m/description.php. MPI-INF-3DHP dataset URL: https://vcai.mpi-inf.mpg.de/3dhp-dataset/.

Received: 27 February 2025; Accepted: 14 August 2025
Published online: 23 August 2025

References

1. Svenstrup, M., Tranberg, S., Andersen, H. J. & Bak, T. Pose estimation and adaptive robot behaviour for human-robot interaction. In *2009 IEEE International Conference on Robotics and Automation*. 3571–3576 (IEEE, 2009).
2. Mehta, D. et al. Vnect: Real-time 3D human pose estimation with a single rgb camera. *ACM Trans. Graph. (TOG)* **36**, 1–14 (2017).
3. Zhang, J. et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* **7**, 46–55 (2022).
4. Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299 (2017).
5. Chen, Y.et al. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7103–7112 (2018).
6. Fang, H.-S., Xie, S., Tai, Y.-W. & Lu, C. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2334–2343 (2017).
7. Newell, A., Yang, K. & Deng, J. Stacked Hourglass Networks for Human Pose Estimation. (Springer, 2016).
8. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5693–5703 (2019).
9. Xiao, B., Wu, H. & Wei, Y. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 466–481 (2018).
10. Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017).
11. Zheng, C. et al. 3D human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11656–11665 (2021).
12. Li, W., Liu, H., Tang, H., Wang, P. & Van Gool, L. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13147–13156 (2022).
13. Zhang, J., Tu, Z., Yang, J., Chen, Y. & Yuan, J. Mixste: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13232–13242 (2022).
14. Li, W. et al. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Trans. Multimed.* **25**, 1282–1293 (2022).
15. Shan, W. et al. P-stmo: Pre-trained spatial temporal many-to-one model for 3D human pose estimation. In *European Conference on Computer Vision*. 461–478 (Springer, 2022).
16. Mehraban, S., Adeli, V. & Taati, B. Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6920–6930 (2024).
17. Zhao, Q., Zheng, C., Liu, M., Wang, P. & Chen, C. Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8877–8886 (2023).
18. Zhu, W. et al. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099 (2023).
19. Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1325–1339 (2013).

20. Li, S. & Chan, A. B. 3D human pose estimation from monocular images with deep convolutional neural network. In *Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1–5, 2014, Revised Selected Papers, Part II 12*. 332–347 (Springer, 2015).
21. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
22. Chen, C.-H. & Ramanan, D. 3D human pose estimation= 2D pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7035–7043 (2017).
23. Zhang, J., Mao, Q., Hu, G., Shen, S. & Wang, C. Neighborhood-enhanced 3D human pose estimation with monocular lidar in long-range outdoor scenes. *Proc. AAAI Conf. Artif. Intell.* **38**, 7169–7177 (2024).
24. Elman, J. L. Finding structure in time. *Cognit. Sci.* **14**, 179–211 (1990).
25. Pavllo, D., Feichtenhofer, C., Grangier, D. & Auli, M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7753–7762 (2019).
26. Carion, N. et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*. 213–229 (Springer, 2020).
27. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
28. Holmquist, K. & Wandt, B. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15977–15987 (2023).
29. Shan, W. et al. Diffusion-based hypotheses generation and joint-level hypotheses aggregation for 3D human pose estimation. In *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
30. Zhou, K., Han, X., Jiang, N., Jia, K. & Lu, J. Hemlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2344–2353 (2019).
31. Zeng, A. et al. Srnet: Improving generalization in 3D human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. 507–523 (Springer, 2020).
32. Chen, T. et al. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 198–209 (2021).
33. Gong, K., Zhang, J. & Feng, J. Poseaug: A differentiable pose augmentation framework for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8575–8584 (2021).
34. Qian, X. et al. Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. arXiv preprint arXiv:2301.07322 (2023).
35. Tang, Z., Qiu, Z., Hao, Y., Hong, R. & Yao, T. 3D human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4790–4799 (2023).
36. Chen, H. et al. Hdformer: High-order directed transformer for 3D human pose estimation. arXiv preprint arXiv:2302.01825 (2023).
37. Mehta, D. et al. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*. 506–516 (IEEE, 2017).
38. Liu, R. et al. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5064–5073 (2020).
39. Paszke, A. et al. Automatic differentiation in pytorch. In *NIPS 2017 Autodiff Workshop* (2017).
40. Duan, H., Zhao, Y., Chen, K., Lin, D. & Dai, B. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2969–2978 (2022).
41. Kingma, D. P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

Acknowledgements

This work was supported in part by the Zhejiang Province Program (2025C01068, 2024C03263, LZ25F020006), and in part by the National Program of China (62172365), and in part by the Macau project: Key technology research and display system development for new personalized controllable dressing dynamic display, and in part by the Ningbo Science and Technology Plan Project (2025Z052, 2025Z062, 2022Z167, 2023Z137), and in part by the Zhejiang Province Regular Undergraduate Universities "14th Five-Year Plan" Higher Education Teaching Reform Project (No. jg20220286).

Author contributions

J.Y.: Methodology, Validation, Visualization, Writing an original draft. J.C. and Z.L.: Investigation, Resources, Supervision, Reviewing and Editing. J.H. and Y.X.: Formal analysis, Reviewing and Editing. Y.L., L.Z. and W.X.: Conceptualization, Reviewing and Editing. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L., J.Y. or Y.X.