Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 1, 2026

Research on a Transformer Algorithm with Mixture-of-Experts for 3D Human Pose Estimation

Zhongzhuo Li

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

Abstract: Monocular 3D Human Pose Estimation Technology has been widely applied in the field of intelligent rehabilitation devices. However, existing Transformer-based methods typically adopt a uniform parameter structure, making it difficult to effectively capture the significant pose feature differences among various action categories. This leads to suboptimal performance when handling complex or unseen actions. In recent years, the Mixture-of-Experts (MoE) architecture has achieved breakthrough progress in large-scale models, offering new insights for human pose estimation. Inspired by DeepSeekMoE, this paper proposes a novel 3D human pose estimation framework, PoseMoE Former, which significantly enhances the model's generalization ability and modeling efficiency across different action categories by integrating both shared expert and routing expert mechanisms. Specifically, our model comprises three core modules: static and dynamic routing mechanisms that enable expert selection; shared expert and routing expert modules that capture general low-level features and category-specific features; and a loss system including the main task loss, dynamic loss, and load balancing loss, which alleviates issues of unbalanced expert utilization. Experiments on Human3.6M demonstrate that PoseMoE Former outperforms mainstream baseline methods, particularly excelling in complex action categories. Ablation studies further confirm the rationality of each module's design, showing not able academic value and practical application potential.

Keywords: 3D Human Pose Estimation, Transformer, Mixture of Experts, Dynamic Routing, Shared Experts.

1. Introduction

To address the limited modeling capability of existing Transformer-based 3D human pose estimation algorithms when dealing with multi-category and diverse actions, this paper proposes a novel PoseMoE Former architecture with category specificity and dynamic adaptability. The model introduces a shared expert and routed expert structure in the feed-forward network of the Transformer Block: the shared expert captures common low-level features across all action categories, while the routed experts learn specialized feature representations tailored to different action categories. Meanwhile, the paper systematically designs two expert routing mechanisms—static and dynamic. The dynamic routing adaptively activates a set of experts at the token granularity through a gating network, and adopts Top-k and Top-p strategies to enhance the model's flexibility in modeling complex and ambiguous actions as well as its generalization ability.

To fully leverage the advantages of the expert mechanism, this paper introduces a complementary loss function system for the dynamic routing mechanism, including main task losses (3D reconstruction error and temporal consistency), dynamic loss (constraining expert selection entropy), and load balancing loss (balancing the activation frequency of each expert). These losses collectively optimize model accuracy and training stability. Extensive experiments on Human3.6M, a mainstream 3D pose dataset, demonstrate that PoseMoE Former not only outperforms all leading baseline models in performance—with particularly significant accuracy improvements on complex action categories—but also exhibits limited growth in parameters and computational complexity, rendering it highly valuable for practical applications.

In addition, the paper systematically conducts ablation experiments to thoroughly analyze the impact of key factors—such as the proportion of shared experts and MoE insertion positions—on model performance and resource consumption. It also verifies the general enhancement effect of the expert mechanism when applied to other backbone networks (e.g., KTPFormer). Overall, this paper innovatively integrates the category-specific expert mechanism with dynamic routing strategies, providing a novel method in the field of 3D human pose estimation that balances high accuracy, strong generalization ability, and excellent engineering efficiency.

2. Design and Optimization of Mixture-Of-Experts Transformer for 3D Human Pose Estimation

2.1. Model Background

This paper proposes for the first time a pose category-specific MoE-Transformer architecture, PoseMoE Former, which decomposes the feed-forward network in Transformer into two modules: Shared Expert and Routed Expert. These two modules are designed to learn common low-level features and category-specific features respectively. During the model design phase, both static and dynamic routing mechanisms are explored to identify the optimal expert scheduling approach.

2.2. Model Architecture

PoseMoE Former is modified based on TCPFormer, inheriting its input processing, 2D-3D feature extraction, Transformer Block stacking, and regression head structure. The innovation is concentrated in the FFN component within the Transformer Block.

The network flowchart is illustrated below:

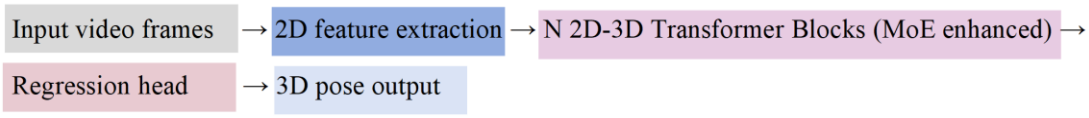

Figure 1. Network flowchart of PoseMoE Former

2.3. Model with Dynamic Routing Mechanism

To address the limitations of static routing, this paper further proposes PoseMoE Former with dynamic routing. A gating network is trained to enable the model to autonomously determine which routing experts should process the input poses (the gating network is jointly pre-trained with the overall model). This improvement transforms routing decisions from discrete categories to continuous confidence scores. Specific implementations include the traditional Top-k mode (k = 2 in this model) and the Top-p dynamic routing mode [1] (p = 0.4 in this model). Comparative experiments verify that the Top-p mode activates fewer experts while achieving slightly better performance. Thus, the Top-p mode is ultimately adopted, with the Top-k mode serving as a control group.

2.3.1. Dynamic Routing Mechanism with Fixed Number of Experts

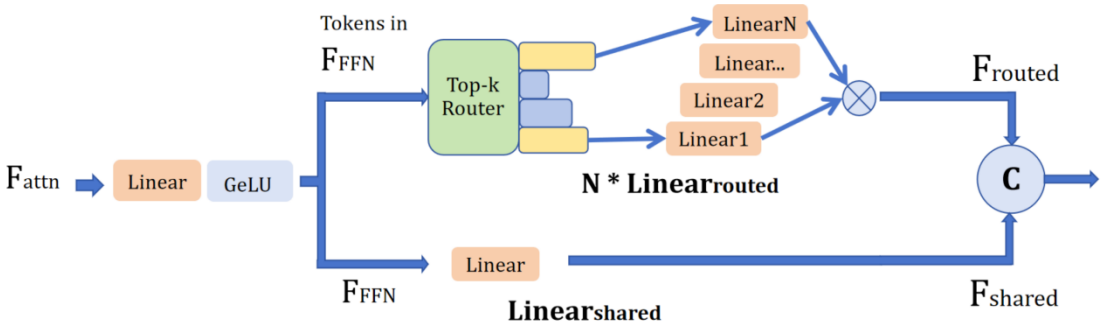

Figure 2. Architecture of PoseMoE Former with Top-k dynamic routing mechanism

As shown in Figure 2, each MoE layer is configured with N routed experts (N=8) and 1 shared expert. Given the output feature F_{attn}, from MHSA, where T is the number of tokens and C is the number of channels. After the first linear transformation of FFN and GeLU activation, we obtain:

$$F_{FFN}=\mathrm{GeLU}(F_{attn}W_1+b_1) \tag{1}$$

Where $W_1\in R^{C\times(eC)}$, and e denotes the FFN expansion factor (e=4 in this model).

Subsequently, F_{FFN} is fed into the shared expert and N routed experts respectively. For the shared expert component, similar to the model with static routing mechanism, it generates features with a dimension of (1-r)C:

$$F_{shared}=F_{FFN}W_{shared}+b_{shared},W_{shared}\in R^{(eC)\times((1-r)C)} \tag{2}$$

For the routed expert component, as illustrated in the upper part of Figure 3–3, for each token in F_{FFN}, a gating network computes the expert selection probability vector P before the token enters the second linear layer of the FFN:

$$P=\mathrm{Softmax}(W_r x^T) \tag{3}$$

Where $W_r\in R^{N\times C}$, C is the token feature dimension, x is the token feature, and $P\in R^N$.

Subsequently, each token passes through the top-k experts with the highest selection probabilities (k=2), and the probabilities of the selected experts are renormalized. The weights corresponding to the other experts are set to zero, indicating that these experts will not be activated. The expert weights are defined as follows:

$$g_i(x)=\begin{cases}\frac{P_i}{\sum_{j\in \mathrm{Top}(P)} P_j}, & i\in \mathrm{TopK}(P)\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

Each expert performs the second linear transformation of FFN on the token as follows:

$$x_{routed}^{(i)}=xW_{routed}^{(i)}+b_{routed}^{(i)},i=1,\ldots,N,W_{routed}^{(c)}\in R^{(eC)\times(rC)} \tag{5}$$

Then, the output of each token is obtained by the weighted average of the routed expert features:

$$x_{routed}=\sum_{i=1}^{N} g_i(x)\cdot x_{routed}^{(i)} \tag{6}$$

The output F_{FFN} of F_{routed} through the routed expert component is formed by concatenating the output x_{routed} of each token.

Finally, the output F_{shared} of F_{FFN} through the shared expert component and the output F_{routed} through the routed expert component are concatenated along the channel dimension to obtain the final output:

$$F_{out}=[F_{shared},F_{routed}] \tag{7}$$

2.3.2. Dynamic Routing Mechanism with Dynamic Number Of Experts

Although the Top-K routing strategy demonstrates good performance, it assumes that each token should be assigned to the same number of experts, ignoring differences in processing difficulty among different input samples. Moreover, since a fixed number of experts are activated in each layer of the Transformer model, this strategy overlooks the differences in representations between layers—specifically, the actual number of experts that need to be activated varies across layers. To address this, this paper adopts the Top-p mechanism, which dynamically determines the activation set based on expert confidence, for improvement.

The Top-p dynamic routing mechanism is illustrated in Figure 3–4. The probability vector P is regarded as the confidence level for selecting each expert, where P_i represents the model's confidence that the i-th expert can adequately process the input token. If the maximum probability in the probability vector is sufficiently high, only the corresponding single expert may need to be activated;

however, if the maximum probability is not large enough, more experts need to be added to improve the reliability of processing the input token. The model gradually increases the number of experts until the cumulative probability of the selected experts reaches or exceeds a predetermined threshold p. To minimize the number of activated experts, the model adds experts in descending order of their selection probabilities. As shown in Figure 3, the upper token has a P_i of a single expert lower than the threshold, so another expert is added to process the token simultaneously; while the lower token has a P_i of a single expert already higher than the threshold, so only that single expert needs to be activated for processing.

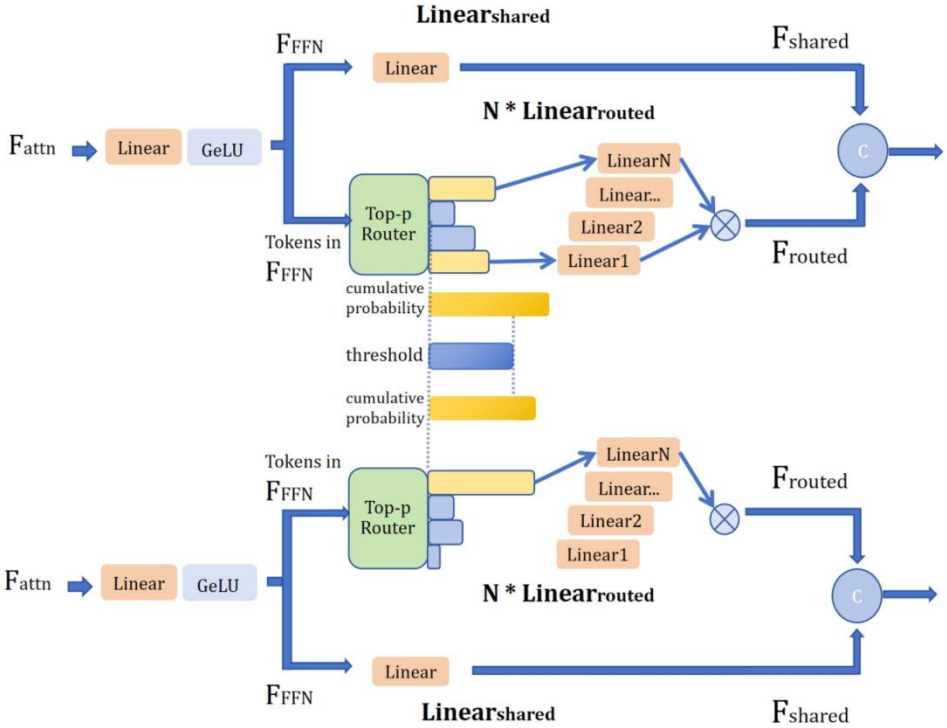

Figure 3. Architecture of PoseMoE Former with Top-p dynamic routing mechanism

Similar to the Top-k dynamic routing mechanism, each MoE layer is configured with N routed experts (N=8) and 1 shared expert. Given the output feature F_{attn} from MHSA, where T is the number of tokens and C is the number of channels, we obtain the following after the first linear transformation of FFN and GeLU activation:

$$F_{FFN}=\mathrm{GeLU}(F_{attn}W_1+b_1) \tag{8}$$

Where $W_1\in R^{C\times(eC)}$, and e denotes the FFN expansion factor (e=4 in this model).

F_{FFN} is fed into the shared expert and N routed experts respectively. For the shared expert component, similar to the model with static routing mechanism, it generates features with a dimension of (1-r)C:

$$F_{shared}=F_{FFN}W_{shared}+b_{shared}, W_{shared}\in R^{(eC)\times((1-r)C)} \tag{9}$$

For the routed expert component, as shown in Figure 3–4, for each token in F_{FFN}, a gating network computes the expert selection probability vector P before the token enters the second linear layer of the FFN:

$$P=\mathrm{Softmax}(W_r x^T) \tag{10}$$

Where $W_r\in R^{N\times C}$, C is the token feature dimension, x is the token feature, and $P\in R^N$.

Different from the Top-k dynamic routing mechanism, we first sort the elements in the probability vector P in descending order to obtain the sorted index list $I=[I_1, I_2, \ldots, I_N]$, where I_j denotes the index of the j-th largest element in the original probability vector after sorting. We then find the smallest expert set S such that the cumulative probability of these experts exceeds the threshold p. The number of experts t to be selected is calculated as:

$$t=\arg\min_{k\in\{1,\ldots,N\}}\left(\sum_{j=1}^{k}P_{I_j}\geq p\right) \tag{11}$$

Here, p is a hyperparameter that controls the confidence level required for the model to stop adding experts. The value of p ranges from 0 to 1, with a higher value resulting in more activated experts. (In our model, p=0.4)

Thus, we determine the set of activated experts S as the experts corresponding to the first t indices after sorting, i.e., $S=\{I_1, I_2, \ldots, I_t\}$. For experts t in S, their weights P_i are retained and then renormalized. For experts not in S, their weights are set to zero. The specific calculation method is as follows:

$$g_i(x)=\begin{cases}\frac{P_i}{\sum_{j\in S}P_j}, & i\in S\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

Subsequently, each expert performs the second linear transformation of FFN on the token as follows:

$$x_{routed}^{(i)}=xW_{routed}^{(i)}+b_{routed}^{(i)}, i=1,\ldots,N, W_{routed}^{(c)}\in R^{(eC)\times(rC)} \tag{13}$$

Then, the output of each token is obtained by the weighted average of the routed expert features:

$$x_{routed}=\sum_{i=1}^{N}g_i(x)\cdot x_{routed}^{(i)} \tag{14}$$

The output F_{routed} of F_{FFN} through the routed expert component is formed by concatenating the output x_{routed} of each token.

Finally, the output F_{FFN} of F_{shared} through the shared

expert component and the output F_{routed} through the routed expert component are concatenated along the channel dimension to obtain the final output:

$$F_{out}=[F_{shared},F_{routed}] \quad (15)$$

The dynamic mechanism can adaptively activate a varying number of experts, enhancing modeling capability and efficiency. Experiments demonstrate that the Top-p scheme achieves better performance.

For both static and dynamic MoE mechanisms, PoseMoE Former only computes a small number of additional routed experts in the second linear layer of the FFN. Compared with the original TCPFormer, it incurs almost no extra parameters or computational cost, rendering it highly valuable for practical applications.

2.4. Loss Function Design

2.4.1. Main Task Loss (3D Pose Estimation + Temporal Consistency)

The main task loss L_{pose} includes the 3D pose reconstruction loss and the temporal consistency loss.

(1) 3D pose L2 loss

We use the L2 loss to minimize the error between the predicted values and the ground truth. Let J be the number of key points, T be the number of frames, and $\hat{Y}_{j,t}$, $Y_{j,t}$ be the predicted and ground truth 3D coordinates of the j-th key point in the t-th frame.

$$L_{3D}=\frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T}\left\|\hat{Y}_{j,t}-Y_{j,t}\right\|_2 \quad (16)$$

(2) Temporal consistency loss

To generate smooth and realistic motions, we introduce the temporal consistency loss proposed by Hossain et al. in 2018 [2]. It is defined as follows:

$$\Delta\hat{Y}_t=\hat{Y}_t-\hat{Y}_{t-1},\Delta Y_t=Y_t-Y_{t-1} \quad (17)$$

Then

$$L_{TC}=\frac{1}{J(T-1)}\sum_{j=1}^{J}\sum_{t=2}^{T}\left\|\Delta\hat{Y}_{j,t}-\Delta Y_{j,t}\right\|_2 \quad (18)$$

The final main task loss is:

$$L_{pose}=L_{3D}+\lambda L_{TC} \quad (19)$$

Where λ is a balance coefficient used to balance positional accuracy and motion smoothness.

2.4.2. Dynamic Loss

The Top-p dynamic routing mechanism has a potential risk: it may assign low confidence to all experts, thereby activating more experts to pursue better performance. Suppose the probability distribution P is uniform and the hyperparameter p is set to 0.5; the model will activate at most half of the experts. This violates the original intention of the MoE framework to efficiently expand model capacity.

To prevent the dynamic routing mechanism from using excessive parameters and losing selectivity for experts in the aforementioned manner, we introduce an additional constraint on the probability distribution P. To make the routing mechanism select only a necessary few experts, we aim to minimize the entropy of P, thereby ensuring each token focuses on as few experts as possible. For this purpose, we design a dynamic loss term that encourages the routing mechanism to select the minimal necessary set of experts, with the formula as follows:

$$L_{dynamic}=-\sum_{i=1}^{N}P_i\cdot\log(P_i) \quad (20)$$

Where P_i is the probability that the token selects the i-th expert

2.4.3. Load Balancing Loss

MoE models typically require distributed training, where different expert modules are deployed on different computing nodes. To avoid situations where resources of some nodes are fully utilized while those of others are underutilized—thus affecting overall training efficiency—we generally aim for a balanced number of tokens processed by each expert. Additionally, existing research (Zuo et al. [3], 2022) has shown that uniform activation of experts in MoE layers helps improve model performance. To achieve load balancing among experts, this work introduces the load balancing loss $L_{balance}$, a method widely adopted in numerous related studies (e.g., Fedus et al. [4], 2022).

$$L_{balance}=N\cdot\sum_{i=1}^{N}f_i\cdot Q_i \quad (21)$$

$$f_i=\frac{1}{M}\sum_{j=1}^{M}1\{e_i\in S^j\},Q_i=\frac{1}{M}\sum_{j=1}^{M}P_i^j$$

Where f_i denotes the proportion of tokens that select expert e_i, and Q_i denotes the proportion of routing probabilities assigned to expert e_i. M is the total number of tokens, S^j is the set of activated experts for the j-th token, and P^j is the probability that the j-th token selects each expert.

2.4.4. Total Loss Function

The total loss of PoseMoE Former is the weighted sum of the main task loss, load balancing loss, and dynamic loss:

$$L=L_{pose}+\alpha L_{balance}+\beta L_{dynamic} \quad (22)$$

Where α and β are hyperparameters used to adjust the contribution of the load balancing loss and the dynamic loss. In the experiments, $\alpha=1\times10^{-2}$ and $\beta=1\times10^{-4}$.

3. Experiment and Result Analysis

3.1. Data Set and Evaluation Index

3.1.1. Human3.6M Data Set

Human3.6M is one of the largest indoor 3D pose estimation data sets, including about 3.6 million image sequences and corresponding 3D human key points. The data were collected by 11 professional actors (6 males and 5 females) performing 15 kinds of daily actions (such as walking, talking, eating, talking on the phone, sitting in a chair, etc.) in the laboratory environment. The collection site is an indoor space of 4m×3m. 3D joint coordinates are recorded by a high-precision optical motion capture system (Vicon), and multi-view video images are obtained by using four synchronous cameras. The dataset provides 3D coordinates and corresponding 2D projection coordinates of 17 joint points, as well as labeling information such as human skeleton topology and camera parameters. Human3.6M is usually divided according to the standard protocol: five subjects (such as S1, S5, S6, S7, S8 and S8) are used as training sets, and the data of the other two subjects (such as S9 S9, S11) are used for test and evaluation. When evaluating, calculate the error between the model prediction and the true value, and you can choose to train/evaluate each action individually or train a single model to evaluate all actions in a unified way.

3.1.2. MPI-INF-3DHP Data Set

MPI-INF-3DHP is a single-person 3D pose data set covering indoor and outdoor scenes, which was proposed by MPI Research Institute and collected by using an unmarked multi-camera motion capture system. This data set contains more than 1.3 million image sequences and corresponding 3D pose annotations. The data came from the pictures taken by 14 synchronous cameras, and recorded 8 kinds of behaviors (such as walking/standing, exercise, sitting posture, squatting/bending, ground movements, sports events, etc.) of 8 subjects (4 males and 4 females) in indoor green screen shed, indoor general environment and outdoor natural scene. Compared with Human3.6M, which is mainly collected in a controlled room, MPI-INF-3HP introduces more kinds of costumes, lighting and backgrounds, and has higher diversity of actions and scenes. MPI-INF-3HP is usually used to test the generalization performance of existing models: it can be trained on its training set and evaluated on the test set, and it is also often used to evaluate cross data sets (for example, the model trained by Human3.6M is directly evaluated on the 3HP test set) to measure the adaptability of the model to unknown scenarios.

3.1.3. Evaluation Indicators

We use commonly used indicators MPJPE, PCK and AUC in the field of 3D human posture estimation to quantitatively evaluate the model. Among them, mpjpe (mean per joint position error) represents the average joint position error, and the average value of Euclidean distance between 3D joint coordinates predicted by the calculation model and the corresponding real coordinates is usually in millimeters. When calculating, the prediction is usually aligned with the root of the real posture (hip joint) to eliminate the influence of the overall translation deviation [5]. The smaller the MPJPE value, the higher the 3D reconstruction accuracy. PCK (percentage of correct key points) is an index to evaluate the hit rate of key points. We use 3D PCK@150mm, that is, statistically predict the ratio of joint points with the real position error within 150 mm. 150mm is about half the size of an adult's head, which is a commonly used threshold in literature [6]. AUC (Area Under Curve) is the area under the curve of PCK changing with the error threshold, and we calculate the AUC percentage of PCK curve within the threshold range of 0 to 150mm [7]. PCK and AUC can more intuitively reflect the performance of the model under different precision requirements: PCK@150mm focuses on measuring the prediction coverage in a large error range, while AUC integrates the performance under various thresholds from strict to loose. Generally speaking, the higher the PCK and AUC values, the higher the accuracy and robustness of the model. In the evaluation, we compare the MPJPE (the lower the better), PCK and AUC (the higher the better) of each model.

3.1.4. Reasons for Using the True Values of Two-Dimensional Key Points as Input

In this paper, all the methods in the experiment use the artificially marked two-dimensional joint position (GT) as input, instead of the actual image or the two-dimensional point predicted by the detector. The purpose of this is to eliminate the influence of two-dimensional attitude detection errors and focus on evaluating the performance of the model from two-dimensional key points to three-dimensional attitude reconstruction. Because the two-dimensional human posture detection itself may produce errors, if the three-dimensional posture is predicted directly from the end-to-end image, it will be difficult to judge whether the errors come from the two-dimensional detection or the three-dimensional reconstruction stage. Using the true values of two-dimensional key points as input can provide ideal and consistent input, ensure that all comparison models work under the same and perfect two-dimensional observation, and thus fairly compare the effects of three-dimensional reconstruction algorithms themselves. As Martinez et al. [8] pointed out, using the true values of two-dimensional key points as input can achieve the theoretical best three-dimensional reconstruction accuracy (the average error of the model is 37.1mm when the true values of two-dimensional key points are input, which is about 30% higher than when the two-dimensional input is detected [9]), so this practice can be regarded as giving the upper limit of the model performance.

3.2. Baseline Model Results and Comparative Analysis

We strictly reproduce six mainstream Transformer-based three-dimensional human pose estimation algorithms (including TCP- Former, MotionAGFormer, KTPFormer, STCFormer, PoseFormerV2, MotionBERT), and complete a comprehensive comparison of the main indicators MPJPE, PCK, AUC on the two public data sets of Human3.6M and MPI-INF-3HP. The results of the repeated experiments are shown in Table 1 on the next page and Table 2 on the second page, respectively, and the error is controlled within 0.3 mm/0.3%, so as to ensure the rigor of the subsequent comparison.

Table 1. Comparison of MPJPE performance of each model in Human 3.6m data set (unit: mm)

Model	T	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	MPJPE
TCPFormer(AAAI2025)	243	15.3	16.1	16.4	14.4	16.0	16.7	16.5	17.2	22.4	21.0	17.1	13.7	14.2	9.6	10.3	15.8
	81	19.5	20.0	20.5	18.6	21.0	24.4	21.2	20.7	27.3	27.3	22.0	18.9	20.0	16.2	15.8	20.9
MotionAGFormer(WACV2024)	243	16.5	17.4	17.7	15.5	17.6	19.6	19.1	18.9	23.1	21.6	18.0	16.8	16.2	11.9	12.6	17.5
	81	20.7	21.5	21.5	19.8	21.7	25.4	23.0	20.9	26.4	28.0	22.7	20.5	21.1	16.2	16.6	21.8
KTPFormer(CVPR2024)	243	19.6	18.6	18.5	18.1	18.7	22.1	20.8	18.3	22.8	22.4	18.8	18.1	18.4	13.9	15.2	19.0
	81	22.5	22.4	21.3	21.4	21.2	25.5	24.2	22.4	24.4	27.5	22.7	21.4	21.7	16.3	17.3	22.2
STCFormer(CVPR2023)	243	20.9	21.9	20.1	20.7	23.5	25.1	23.7	19.4	27.9	26.2	21.7	20.7	19.6	14.3	15.2	21.4
	81	26.1	26.0	22.8	24.1	24.7	27.7	27.7	23.2	30.3	31.7	25.2	24.8	23.9	18.5	19.7	25.1
PoseFormerV2(CVPR2023)	243	22.2	22.5	21.9	21.2	24.0	26.3	25.0	20.8	29.1	27.8	22.9	21.7	20.6	15.4	16.1	22.5
	81	27.6	27.3	24.2	25.3	25.9	28.8	28.7	24.5	31.8	33.0	27.0	25.7	24.8	19.5	20.3	26.3
MotionBERT(ICCV2023)	243	16.2	17.5	17.2	14.7	17.1	18.9	18.8	18.7	22.2	22.1	17.5	17.2	16.3	10.7	11.5	17.1
	81	19.7	20.5	20.7	18.0	21.4	25.4	22.9	21.3	26.5	27.5	21.7	21.5	21.4	16.2	16.3	21.4

Table 2. Performance Comparison of MPJPE, PCK and AUC of each model in MPI-INF-3D HP data set

Model	T	MPJPE (mm)	PCK (%)	AUC (%)
TCPFormer (AAAI2025)	81	15.1	98.8	87.4
	27	18.0	98.5	86.3
MotionAGFormer (WACV2024)	81	16.3	98.1	85.2
	27	19.3	98.1	83.2
KTPFormer (CVPR2024)	81	16.9	98.8	85.8
	27	19.4	98.8	84.2
STCFormer (CVPR2023)	81	23.2	98.6	83.8
	27	24.3	98.4	83.3
PoseFormerV2 (CVPR2023)	81	27.8	97.8	82.0
	27	31.0	97.5	81.3
MotionBERT (ICCV2023)	81	16.0	98.5	86.8
	27	18.7	98.4	85.4

Repetition experiments show that TCPFormer performs best in all baseline, and the accuracy of time window T = 243 is significantly better than T = 81, which verifies the positive effect of long-time series modeling on three-dimensional attitude estimation. In addition, it is found that simple categories such as "Walk/WalkT/WalkD" have the lowest error (9.6~ 14.2mm under TCPFormer T = 243), while complex categories such as "Sit/SitDown/Photo" have significantly increased error (> 20mm). This shows that even with the current optimal baseline, there is still a bottleneck in the modeling ability for complex categories, and also demonstrates the practical necessity of our class-specific routing mechanism from the side.

3.3. Performance Evaluation of The Proposed Model

3.3.1. Experimental Setup

We have implemented a series of PoseMoE Former models on PyTorch platform. The main parameters are set as follows: the number of modules N = 16, the number of attention heads H = 8, the hidden feature dimension C = 128, the time dimension L = T/3 of the hidden attitude agent, and the agent initialization distribution d adopts Gaussian distribution. During training, each mini-batch contains 16 sequences, and 90 epoch are trained with AdamW optimizer, with weight attenuation of 0.01, initial learning rate of 5e-4, exponential attenuation and attenuation factor of 0.99.

3.3.2. Main Results

On the Human3.6M data set, PoseMoE Former takes 2D ground truth as input, and adopts the evaluation process consistent with the mainstream methods. The results are shown in Table 3 on the next page, where V2 uses the Top-p strategy. The results show that:

PoseMoE Former V2 (Dynamic Top-p Routing) performed the best, with an average MPJPE of 14.7mm, which was significantly higher than TCPFormer (15.5mm) and MotionBERT (17.8mm). PoseMoE Former V1 (static routing) MPJPE is 15.2mm, which is also superior to all other baseline models, fully demonstrating the effectiveness of "category expert mechanism". However, although static routing is also effective, it is slightly less flexible when it comes to actions that are difficult to strictly classify, resulting in its performance slightly lower than V2. PoseMoE Former series achieved a lead of 0.8 mm (V2) and 0.3mm(V1) compared with SOTA algorithm (TCP former). Further analysis of various categories shows that PoseMoE Former V2 is obviously superior to TCPFormer and other strong baselines in complex actions such as "SitDown" and "Photo", which shows that the modeling ability of "category-specific+dynamic expert mechanism" for complex actions is significantly enhanced.

Table 3. Comparison of mainstream methods of Human 3.6m data set and MPJPE(GT) of PoseMoE Former series (unit: mm)

Model	T	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	Avg.
MixSTE (CVPR2022)	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
P-STMO (ECCV2022)	243	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
STCFormer (CVPR2023)	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
GLA-GCN (ICCV2023)	243	20.1	21.2	20.0	19.6	21.5	26.7	23.3	19.8	27.0	29.4	20.8	20.1	19.2	12.8	13.8	21.0
MotionBERT (ICCV2023)	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
KTPFormer(CVPR2024)	243	19.6	18.6	18.5	18.1	18.7	22.1	20.8	18.3	22.8	22.4	18.8	18.1	18.4	13.9	15.2	19.0
TCPFormer (AAAI2025)	243	15.0	15.9	16.1	14.2	15.7	16.4	16.2	16.9	22.1	20.6	16.7	13.3	13.8	9.2	9.9	15.5
PoseMoE Former V1(Ours)	243	14.8	15.6	15.7	14.2	15.6	16.0	15.8	16.6	21.5	20.1	16.5	13.2	13.4	9.0	9.8	15.2
PoseMoE Former V2(Ours)	**243**	**14.2**	**15.0**	**15.3**	**13.7**	**15.1**	**15.5**	**15.2**	**16.0**	**21.1**	**19.7**	**16.1**	**12.7**	**13.2**	**8.8**	**9.6**	**14.7**

In addition, we have also done qualitative experiments on unusual complex action videos in the real environment, as shown in Figure 4. As can be seen from the figure, the performance of our proposed PoseMoE Former is also significantly better than that of the baseline model TCPFormer: V1 model is better than TCP former in the head, arms and other parts; The V2 model is obviously closer to the input video in the upper torso part with serious occlusion, which shows that the PoseMoE Former proposed by us has significantly enhanced the modeling ability of complex actions.

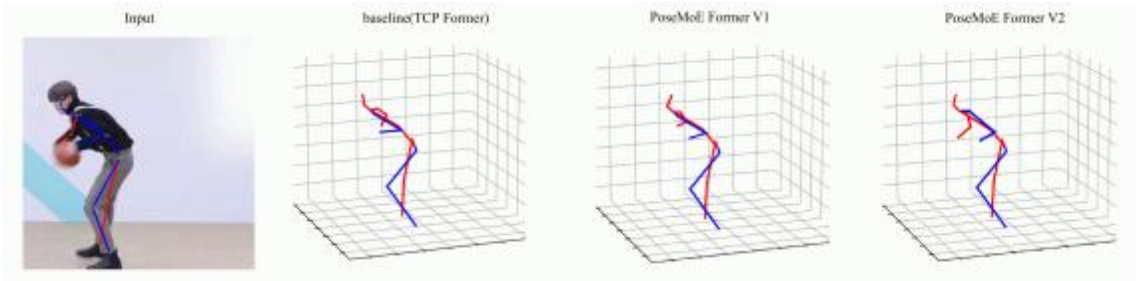

Figure 4. Qualitative comparison between Posemoe former and TCPFormer in real-world video

3.3.3. Comparison Between the Number of Fixed Experts and The Number of Dynamic Experts

In order to further explore the influence of different dynamic routing strategies on the performance and efficiency of the model, the Top-k(k=2) and Top-p(p=0.4) strategies are compared systematically, and the results are shown in Table 4.

Table 4. Comparison of performance and efficiency of PoseMoE Former model under different routing strategies

Strategy	MPJPE (GT) (mm)	Parameter Param (M)	FLOPs (G)
top-k (k=2)	14.8	74.2	252
top-p (p=0.4)	14.7	74.2	245

Under the main index MPJPE(GT) of Human3.6M, the current optimal performance of the Top-p strategy is 14.7mm, and the FLOPs is 245G, which is about 3% lower than that of the Top-k scheme (252G). The reason is that Top-k always forces each token to activate a fixed number of experts, ignoring the actual difficulty or category distribution of tokens, resulting in a waste of some token resources; The Top-p mechanism can dynamically adjust the number of experts according to the confidence of the gated network. Simple or well-informed tokens activate fewer experts, while complex or fuzzy tokens activate more, which makes the resource allocation more reasonable and efficient, making the average number of activated experts slightly lower than 2 when p=0.4, so the overall FLOPs decreases. Although Top-p reduces the amount of calculation, the performance of the model has not decreased, but has improved slightly. This shows that the Top-p mechanism can capture the specific modeling requirements of each token more accurately, reduce unnecessary redundant feature mixing, and thus improve the adaptability and generalization of diversified actions.

3.4. Ablation Experiment

3.4.1. Influence of the Proportion of Shared Experts on Performance

This part evaluates the influence of the proportion of shared experts in PoseMoE Former on performance and resource expenditure, and the results are shown in Table 5.

Table 5. Influence of the proportion of shared experts in Posemoe former on performance and resource overhead (Top-p routing)

Model	Proportion of shared experts	MPJPE (GT) (mm)	Parameter Param (M)	FLOPs (G)
TCPFormer	~1	15.5	35.1	218
PoseMoE Former/ have sharing experts	0.67	15.1	61.6	236
PoseMoE Former/ have sharing experts	0.5	14.8	74.2	245
PoseMoE Former/ have sharing experts	0.33	14.9	87.8	255
PoseMoE Former/ have sharing experts	~0	15.1	115.1	273

Note: Top-p strategy is used here.

The results of ablation experiments show that the reasonable proportion of sharing experts and routing experts in PoseMoE Former is very important to improve performance. Compared with TCPFormer which only adopts fully shared parameters, the introduction of sharing and routing expert decomposition can significantly reduce MPJPE (15.5mm→ 14.8mm). When the proportion of shared experts is set to 0.5, the model achieves the optimal balance among performance, parameter quantity and reasoning efficiency. Further reducing the sharing ratio, although the model parameters and FLOPs increase, the performance decreases, which shows that the sparse expert allocation caused by over-customization is not conducive to generalization. Although the number of parameters of extreme full routing structure has increased sharply, it has not brought about performance improvement, but has aggravated the consumption of model resources. It can be seen that the reasonable proportion of sharing and routing experts takes into account both the bottom universality (sharing experts) and the high-level category/action specificity (routing experts), which can maximize efficiency and model performance.

3.4.2. Insertion Position of Hybrid Expert Structure in Feedforward Neural Network

The ablation experiment further explored the influence of the insertion position of MoE expert mechanism in FFN structure on the model performance and resource consumption, and the results are shown in Table 6. The experimental results show that the best performance can be obtained only by introducing the expert mechanism into the second linear layer of FFN (MPJPE=14.7mm), and the parameters and calculation amount are much lower than that of the full FFN-MoE scheme (74.2M/245G vs 115.3M/273G). The introduction of MoE into all linear layers of FFN did not improve the visible performance, but significantly increased the resource burden. This is because the expert mechanism is only introduced in the second linear layer, so that experts can participate in the final feature compression and fusion. This layer has a greater influence on the network expression ability, and it is also the most suitable to realize the specific modeling of categories/actions. Only the second layer MoE scheme retains all the advantages of expert mechanism in improving the model capability, but greatly reduces the model

complexity and calculation cost, making the model easier to deploy and expand. Therefore, PoseMoE Former's choice of "only the second linear layer MoE" in design is an optimal strategy with both high efficiency and high performance.

Table 6. Comparison of performance and efficiency of MOE structure at different insertion positions in FFN

MoE insertion position	MPJPE (GT) (mm)	Parameter Param (M)	FLOPs (G)
Full FFN-MoE	14.7	115.3	273
Only the second linear layer -MoE	14.7	74.2	245

3.4.3. Model Universality Experiment

In order to verify the universality of PoseMoE Former, the dynamic expert routing mechanism is integrated into KTPFormer backbone, and the results are shown in Table 7 on the following page. The results show that the average MPJPE is reduced from 19.0mm to 18.1mm and the largest category is increased by 1.1mm after KTPFormer integrates PoseMoE Former V2. The experiment shows that the expert decomposition structure proposed by PoseMoE Former has universal enhancement ability to the existing Transformer-based 3D HPE model, and can effectively improve the accuracy in complex and diverse action scenes.

Table 7. Experimental results of generality of integrating Posemoe Former V2 into KTPFormer (MPJPE, unit: mm)

MODEL	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	Avg.
KTPFormer	19.6	18.6	18.5	18.1	18.7	22.1	20.8	18.3	22.8	22.4	18.8	18.1	18.4	13.9	15.2	19.0
KTPFormer Integrating the Proposed Method	**18.8**	**17.8**	**17.7**	**17.3**	**17.6**	**21.1**	**19.8**	**17.5**	**21.7**	**21.3**	**18.0**	**17.5**	**17.8**	**13.5**	**14.8**	**18.1**

4. Conclusion

With the increasing demand for intelligent physiotherapy equipment, the research on high-precision three-dimensional human posture estimation has gradually become a hot spot. This paper focuses on the task of monocular three-dimensional human pose estimation, analyzes the limitations and technical challenges of existing methods in detail, and innovatively introduces a hybrid expert mechanism to significantly improve the generalization ability of the model and the modeling efficiency of multi-action categories. Specifically, the main contents and contributions of this paper are summarized as follows:

Firstly, this paper systematically combs the development of monocular 3D human pose estimation method and MoE model, especially the breakthrough of Transformer architecture in this field and the extensive application and structural innovation of MoE model in large-scale deep learning, and combs the current research trends. In the part of theory and background, this paper introduces the concept and structure of Transformer model and MoE, and then deeply analyzes the advantages and disadvantages of different technical routes and different theoretical bases for 3D human posture estimation, and determines the mainstream research content. And technical defects. At the same time, the advantages and disadvantages of the existing Transformer model in dealing with multi-class pose features are compared, especially the traditional unified parameter method is difficult to fully capture the differences of different action categories in pose space.

In the related work part, this paper analyzes the application status of posture categories in the field of human posture estimation in detail, and points out that although existing methods such as ActionPrompt try to improve the performance of posture estimation by using action category information, they still have shortcomings in dealing with complex and unknown action sequences. In addition, the performance and limitations of mainstream baseline models such as TCPFormer are systematically compared. Then, this paper discusses the shared expert mechanism and Top-p dynamic routing mechanism from DeepSeekMoE in detail, thus expounding the motivation of putting forward PoseMoE Former in this paper, that is, to further strengthen the distinguishing ability of attitude category characteristics by introducing expert mechanism.

Based on the above background, this paper puts forward the PoseMoE Former algorithm, which integrates the sharing expert module and the routing expert module into the feedforward network of Transformer to capture the general features and the specific features of action categories respectively, which greatly improves the flexible modeling ability of the model for multi-category actions. Two expert scheduling mechanisms, static and dynamic, are designed: static routing enables specific expert sub-models for predefined action categories, while dynamic routing selects experts in real time through gated networks, which improves the adaptability of the model to complex or unknown action sequences.

Then, in order to effectively train the above-mentioned expert system, this paper designs a matching loss function system, including main task loss (3D reconstruction error and time sequence consistency), dynamic loss (limiting experts' selection entropy) and load balancing loss (ensuring the balanced utilization of experts), and comprehensively optimizes the performance and training stability of the model. Experiments on the mainstream human posture data set Human3.6M show that the proposed PoseMoE Former model is significantly superior to the existing mainstream baseline algorithm in the estimation accuracy of complex motion categories, which further verifies the practical application value of this algorithm.

In addition, through detailed ablation experiments, this paper discusses the influence of key design factors such as the proportion of shared experts and the insertion position of MoE module on the model performance and resource consumption, and verifies that the expert mechanism can enhance the universality of other Transformer- based three-dimensional human posture estimation methods.

Specifically, the innovations and contributions of this paper can be summarized as follows:

(1) A PoseMoE Former for 3D human pose estimation is innovatively proposed, which integrates sharing experts and routing experts, effectively solving the problem that features are difficult to distinguish between different action categories, and designing static and dynamic expert routing mechanisms to improve the accuracy and generalization ability of the model. (2) A special loss function system, including dynamic loss and load balance loss, is constructed, which solves the

problems of expert imbalance and routing instability in MoE architecture training and improves the robustness of model training. (3) Experiments show that the proposed PoseMoE Former has achieved higher accuracy than the current mainstream methods in human posture estimation tasks, especially in complex motion categories, and has certain academic and engineering value.

To sum up, the method proposed in this paper is not only innovative in theory, but also proves its application potential in practical systems through experiments, which provides an effective solution for the research in the field of monocular three-dimensional human posture estimation.

References

[1] Huang, Q., An, Z., & Zhuang, N., et al. (2024). Harder tasks need more experts. arXiv preprint arXiv:2403.07652.

[2] Hossain, M. R. I., & Little, J. J. (2018). Exploiting temporal information for 3D human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 68–84).

[3] Zuo, S., Zhang, Q., Liang, C., et al. (2022). Moebert: From BERT to mixture-of-experts via importance-guided adaptation. arXiv preprint arXiv:2204.07675.

[4] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(1), 1–39. https://jmlr.org/papers/v23/21-1399.html

[5] Peng, J., Zhou, Y., & Mok, P. Y. (2024). KTPFormer: Kinematics and trajectory prior knowledge enhanced transformer for 3D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Mehta, D., Rhodin, H., Casas, D., et al. (2017). Monocular 3D human pose estimation in the wild using improved CNN supervision. In Proceedings of the International Conference on 3D Vision (3DV) (pp. 506–516). IEEE. https://ieeexplore.ieee.org/document/8283659. https://doi.org/10.1109/3DV.2017.00064

[7] Mehta, D., Rhodin, H., Casas, D., et al. (2017). Monocular 3D human pose estimation in the wild using improved CNN supervision. In Proceedings of the International Conference on 3D Vision (3DV) (pp. 506–516). IEEE. https://ieeexplore.ieee.org/document/8283659. https://doi.org/10.1109/3DV.2017.00064

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

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