Dual Attention Poser: Dual Path Body Tracking Based on Attention

Xinhan Di[2]*, Xiaokun Dai[1,3]*, Xinkang Zhang[1,3], Xinrong Chen[1,3]†
[1]Academy for engineering&technology, Fudan Univerisy
[2]Deepearthgo
[3]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Fudan University
deepearthgo@gmail.com, {xkdai21,xinkangzhang21}@m.fudan.edu.cn, chenxinrong@fudan.edu.cn

Abstract

Currently, mixed reality head-mounted displays tracking the full body of users is an important human-computer interaction mode through the pose of the head and the hands. Unfortunately, users' virtual representation and experience is limited due to high reconstruction error when simple transformer network architecture is applied. In this paper, we present a novel model, named Dual Attention Poser, which can learn the whole body reconstruction at a high accuracy. The proposed model consists of three key modules. Among them, dual-path attention encoder is designed to extract feature of the sparse signals. Cross attention mixer module enable the fusion of representation in the double path. Attention-gated-mlp decoder is applied to decode the hidden feature from the sparse input through attention gate. Test results on the AMASS dataset show that Dual Attention Poser can reduce the error by up to 18.2% in comparison with the state-of-the-art results.

1. Introduction

Since the popularity of virtual reality(VR) and augmented reality(AR), interaction in these environments is in great demand. Firstly, hand tracking is fully developed on VR/AR devices for common applications such as virtual reality games, remote medical control and employment of robots. It is applied to the control of virtual and real objects using hands directly and the experience of the real interaction between objects and human. Secondly, human body reconstruction is of wide application for social demand in the virtual reality and augmented reality. For example, face-to-face communication in digital games and remote meetings plays a key role in social activities in the virtual 3D environment. Therefore, body reconstruction is in great demand for rich experience on faces, hands, arms, and legs. As collaborative interactive with the full body tracking is demonstrated to exceed manual first-person tasks, the embodiment of full avatar representation is well explored through a variety of hardware equipment.

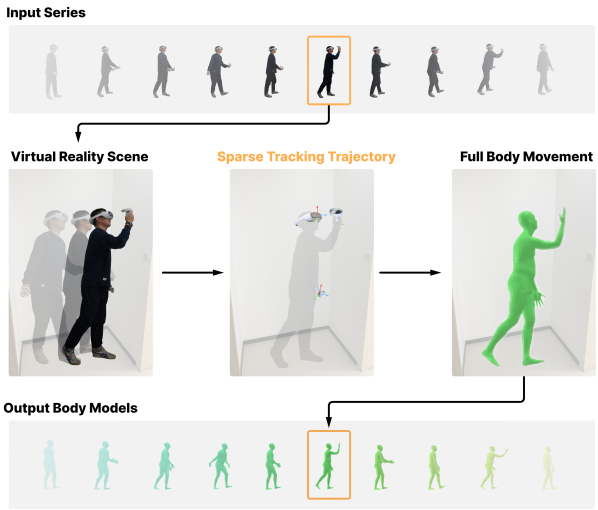

Figure 1. Attention-based Dual Attention Poser can generate a full-body pose over 22 joints with the input of the positions and orientations of one headset and two hands only. Compared with the state-of-the-art methods, our method reaches the highest prediction accuracy, which significantly enhances the performance and immersion of Mixed Reality.

There are a variety of hardware equipment which are designed for body tracking based on sparse input available. Additional trackers are employed on user's body to render the body of an avatar, such as 6D pelvis tracker [38], body-worn inertial sensors [13]. The motion diversity of subjects is extended to give more sparse input for rendering the body and supporting the representation of the body to be more natural and accurate. However, for the common hardware of virtual reality, an accurate full body reconstruction based on the sparse signals from two controllers and one headset is far from satisfactory accuracy. Extra dense motion capture

[*] These authers contributed equally to this work.
[†] Corresponding author

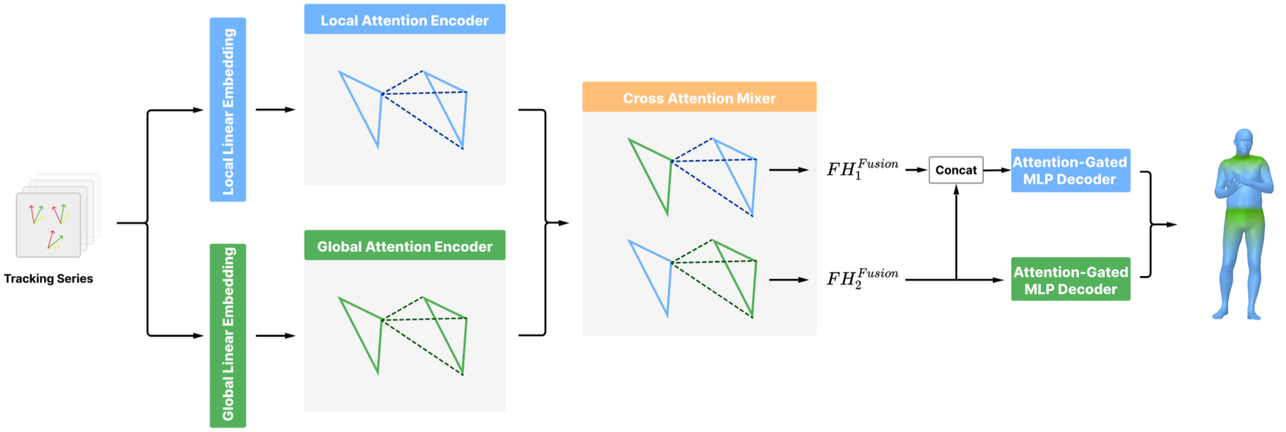

Figure 2. End to end full body reconstruction from sparse input. The network proposed consists of three key modules, including dual path attention encoder, cross attention mixer, and attention-gated-mlp decoder.

system is expensive and common consumers are unwilling to afford it. Therefore, the methods on the post-process of body joints are well explored to handle failures of body reconstruction.

Among the existing works, inverse kinematics(IK) based on human biological motion is usually applied for the estimation of full-body poses. Inverse kinematics(IK) is well explored in the field of robotics to calculate the joint parameters, which is widely used in the application [22,26,29,36] of robotics and computer animation [9, 10, 25, 30, 43] via iterative optimization. The IK based optimization on this numerical iteration is lack of generative ability and the optimization is time-consuming. For the limits, the neural networks are applied to calculate the desired end-effect location, which can speed up the computation. However, the learning based IK methods only refine the position of joints a little.

Besides, in order to improve the accuracy of body reconstruction, a general method is to increase the number of sparse sensors. A variety of inertial measurement unit(IMU) sensors are applied to reconstruct the legs of the body, where LSTM [41] are applied in the models to process the signals of the IMU sensors in real-time. However, users are required to wear more sensors, which is inconvenient. Meanwhile, it's not friendly to wear extra equipment on legs and arms to attend virtual social meeting or other virtual activities. In addition, the reconstruction accuracy of the whole body is not fully satisfied when users play virtual games or attend virtual social activities. Furthermore, the extra equipment for the tracking of the whole body is not easily carried on and the users have to use them at a particular location such as home or VR room.

Then, transformer based methods [14] are applied to increase the accuracy of the body, where attention modules provide a better presentation of sparse input from sensors. In [14], sparse input are decoded into features of the two paths and fed into the Forward-Kinematics Module, in which they are modeled separately to be the joints of the body limbs and the rotation of the whole body, as shown in the blue and green parts on Figure2. However, the local and the global feature are not be separated before decoder, and both the mutual information and the relationship between the local body pose and the global orient is not be used.

Therefore, we proposed a novel model with the combination of separate encoders and fusion modules which learn the mutual benefits and the relationship between the local and the global paths, and conduct ablation studies on different variants of the proposed multi-task model. The proposed model is designed consisting of three key modules. One module, dual path transformer encoder, is applied to extract the local and the global information in the separated path. Cross attention module can fuse the representation of the two paths. Another module, attention-gated-mlp decoder, is used to generate 6 DoF pose of landmarks of the body and reconstruct the whole body.

The method proposed is tested on AMASS dataset. The results show that the method achieves the higher accuracy of full body reconstruction and the error is reduced by up to 18.2%, in comparison with the state-of-the-art methods including transformer based methods and non-transformer methods. Besides, quality evaluation is conducted to demonstrate the effectiveness of our proposed model.

2. Related Work

The following works of full body pose estimation and reconstruction, vision transformers, and multilayer perceptron networks are the most related work.

2.1. Full Body Pose Estimation and Reconstruction

Among the previous works on body pose estimation, the 3D skeleton is an effective form for the representation of human body [34]. The prediction of 3D joints is produced through the application of joints positions [23] and volumetric heat-maps [28]. Then, in order to encode pose and shape parameters of the human body, SMPL known as the skinned multi-person linear model is built for the generation of triangular faces of human body. Afterwards, the reconstruction of human body is built with both optimization and regression methods. For example, the optimization of the parameters of SCAPE [3] model is used for 2D keypoints annotation. A CNN network is used to regress the parameters from the silhouettes and 2D joint heatmaps [27]. However, the input of images are not practical for the full body reconstruction from virtual reality glass/augmented reality glass. The reason is that it's hard to capture the down body part such as legs for the cameras of these glasses. Therefore, body reconstruction for these glasses is then predicted from sparse sensors. 6 body-worn inertial sensors [13, 35, 39, 40] are used as the sparse sensors. They are commonly distributed over head, arms, pelvis, and legs. Firstly, a KNN-based method [1] is applied to interpolate poses from a smaller dataset with only specific motion. Secondly, a GRU network is trained for the prediction of the lower-body pose from the sparse input and the calculation of upper-body pose through an IK solver [38]. However, these techniques of full body reconstruction from both dense and sparse input are not effective and have high errors. They are not practical for the application on virtual reality devices.

2.2. Vision Transformer

Due to the success of transformer in the vision tasks, attention modules are widely used in the vision tasks and the performance is improved significantly in the following tasks particularly, including image classification [7, 8, 19], detection [4, 31, 45], image restoration [16, 37, 42]. Then, transformer encoders and decoders are applied to represent the vertex of human pose and mesh from a single image [17]. Similarly, the body joint correlations and temporal dependencies are learned through transformers [2, 44]. Afterwards, the prediction of multiple pose hypotheses is produced from monocular videos through learning spatio-temporal representation of body poses [15]. However, images from HMD devices are not able to capture the whole body, it's hard to reconstruct the whole body on HMD devices. Therefore, instead of body pose prediction from images or videos, transformer encoders are applied to represent the body reconstruction from sparse inputs [14]. But the architecture of the attention encoder is too simple to achieve low reconstruction errors, so the reconstruction from this model is not practical.

2.3. Multi-layer Perceptron Networks

Multi-layer perceptron(MLP) networks are early explored for image classification and other simple tasks. Recently, a pure MLP network architecture is proposed in which all layers consist of MLP layers, and the architecture is specially designed for vision tasks [32]. Then, the graph convolution layer and mlp layer [12] are fused for the vision tasks, such as image classification and object detection. Besides, dense prediction of image tasks [5] is learned through the cycle-like structure of network architecture consisting of mlp layers. Afterwards, a global filter is applied for the effective representation of image features in the tasks of image classification. Finally, the mlp architecture, the attention modules, and a dynamic gate are fused to better represent connectivity of sparese sensors form the two paths. However, these new MLP architectures cannot represent the complexity of full body reconstruction from sparse input of HMD devices. Therefore, the application of the novel mlp architectures as a module can be a promising work.

3. Method

In this paper, a dual-path cross attention model is proposed for full body reconstruction from the sparse input. The proposed network consists of three key modules, which are dual-path attention encoder, cross-attention mixer, and attention-gated-mlp decoder.

3.1. Problem Formulation

The Problem of the full body reconstruction from sparse input is defined as the following. Given 6D representation of the sparse input, which is consisted of $p^{1\times3}$ representing the global positions in Cartesian coordinates and orientations in axis-angle representation $\theta^{1\times3}$, the model proposed is to learn a mapping f which reconstructs the joints of the full body. The mapping function f is represented as the following:

$$\mathbf{U}_{1:j}^{1:h} = f(\{p, \theta\}_{1:j}^{1:n}) \tag{1}$$

where n represents the number of the sparse input, h represents the number of the joints of the full-body skeleton, j represents the number of observed frames from the past. $\mathbf{U} \in SE(3)$ is the body joint pose which is represented by $\mathbf{U} = \{p, \theta\}$.

Similarly with avatarposer [14], the value of h is set as 22, and the joints are divided into a root joint and 21 local joints to represent the global orient and the other body pose. Then, the representation and animation of SMPL [20] model is applied. In our model, dual path encoder-decoder was utilized to represent the global orient and the other body pose separately. In the following, they are referred to as global and local paths, respectively. Besides, at each time

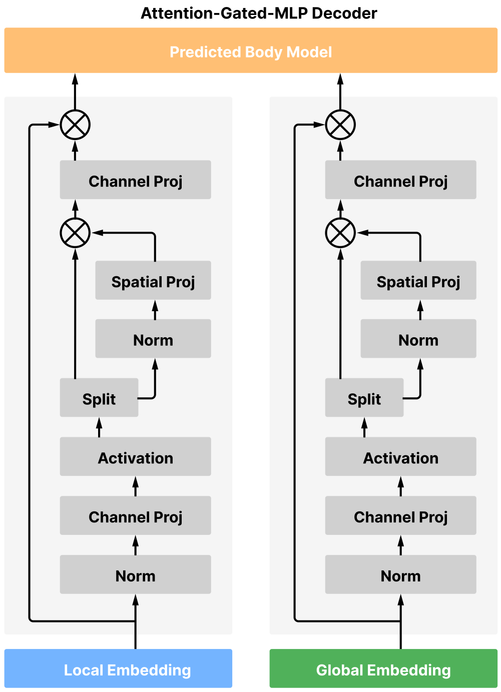

Figure 3. The structure of the attention-gated-MLP Decoder.

step t, the following three features are used. Firstly, a liner velocity v is given by backward finite difference as follows:

$$v_t = p_t - p_{t-1} \tag{2}$$

Secondly, the angular velocity Ω_t is defined as the following:

$$\Omega_t = \mathbf{R}_{t-1}^{-1}\mathbf{R}_t \tag{3}$$

where $R^{3\times3}$ represents the orientation matrix of the sparse input [14], it's calculcated after the conversion [14, 44], the axis-angle representation $\theta^{1\times3}$ is converted to the rotation matrix $R^{3\times3}$.

Thirdly, the last row of R is discarded to get the 6D rotation representation w_t. Then, at each time step t, the feature of each sparse input contains four feature vectors including p_t, v_t, θ_t, w_t. As the number of the sensor is set to 3, the feature of all sparse input is set as the following:

$$\mathbf{X}_t^S = [p_t^1, v_t^1, \theta_t^1, w_t^1, ..., p_t^3, v_t^3, \theta_t^3, w_t^3] \tag{4}$$

Therefore, the number of features for the input $\mathbf{X}$ is 54, and the output dimension at each time step is 132.

3.2. Dual Attention Encoder

For the global path of representation through attention, taking the sparse input $\mathbf{X}$, weights of three projection $W_{Q_1}, W_{K_1}, W_{V_1}$ are calculated to produce the matrices Q_1, K_1, V_1. Then, the attention encoder of the global path is calculated as the following:

$$attention(Q_1, K_1, V_1) = softmax(\frac{Q_1 K_1^T}{d_1})V_1 \tag{5}$$

where d_1 represents the scale factor.

Similarly, for the local path of representation through attention, after the sparse input $\mathbf{X}$ taken, weights of three projection $W_{Q_2}, W_{K_2}, W_{V_2}$ are calculated to produce the matrices Q_2, K_2, V_2. Then, the attention encoder of the local path is calculated as the following:

$$attention(Q_2, K_2, V_2) = softmax(\frac{Q_2 K_2^T}{d_2})V_2 \tag{6}$$

where d_2 represents the scale factor.

As Figure 2 shown, the two paths are calculated separately through the application of multiple layers of attention modules.

3.3. Cross Attention Mixer

Then, a fusion module is calculated through cross attention mechanism. It fuses the hidden features from the local transformer encoder and the global transformer encoder (Figure 2). Let Z_1 represent the hidden features from the local transformer, Z_2 represent the hidden features from the global transformer, the cross attention mixer works to achieve the fusion of the two sparse representation Z_1 and Z_2. The cross attention fusion is calculated as the following:

$$Y_1 = Fusion_1(Z_1, Z_2), \quad Y_2 = Fusion_2(Z_2, Z_1), \tag{7}$$

$$FH^{1\to2} = softmax\left(\frac{f(Q_2)f(K_1)^T}{d_1}\right)f(V_1), \tag{8}$$

$$FH^{2\to1} = softmax\left(\frac{f(Q_1)f(K_2)^T}{d_2}\right)f(V_2), \tag{9}$$

where, Z_1 and Z_2 are the representations of the local path and the global path before the fusion, multi-head self-attention (MHSA) module is applied to obtain the query, key, and value features of the local feature Z_1, and the values are indicated by Q_1, K_1, V_1. Similarly, MHSA module is applied to obtain the query, key, and value features of the global feature Z_2, and the values are represented as Q_2, K_2, V_2. T represents the matrix transpose. $FH^{1\to2}$ and

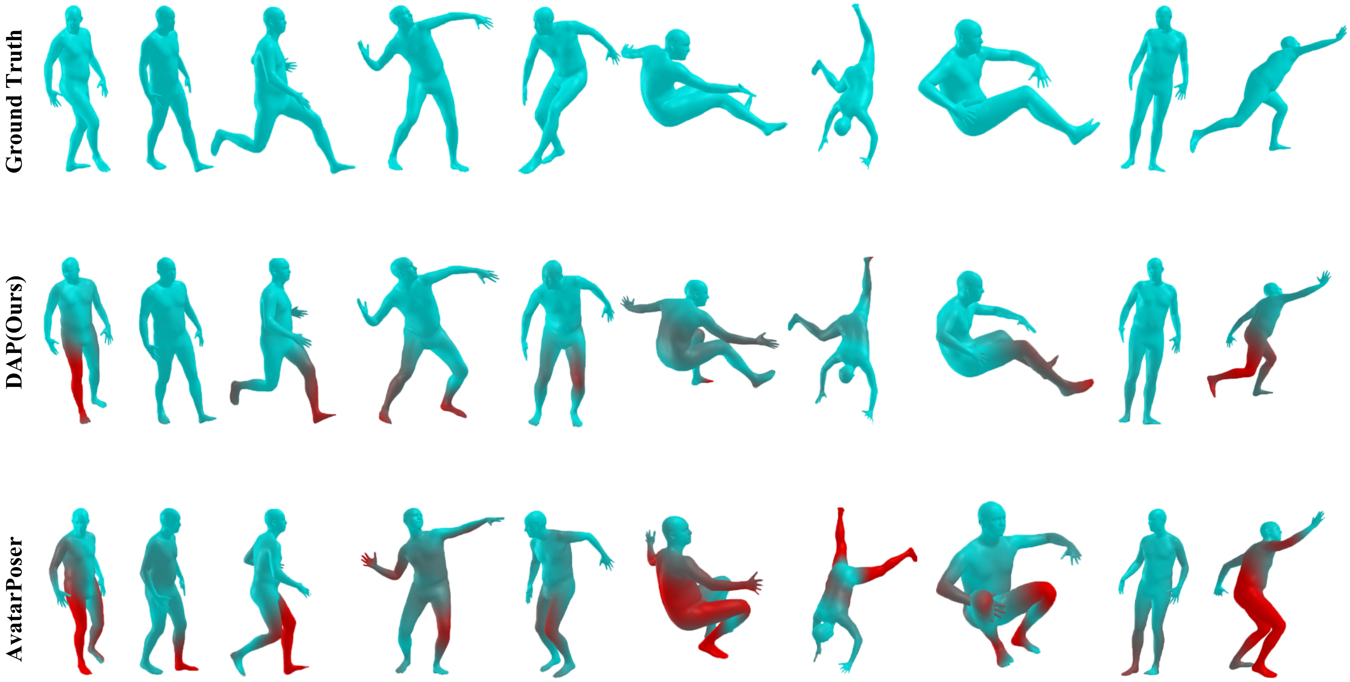

Figure 4. Visual comparison between AvatarPoser [14] and our method. The extent of the error is indicated by the intensity in red. The first row represents the ground truth of reconstruction body. The second row represents the reconstruction results of our models. The third row represents the results of the state-of-the-art model Avatarposer [14].

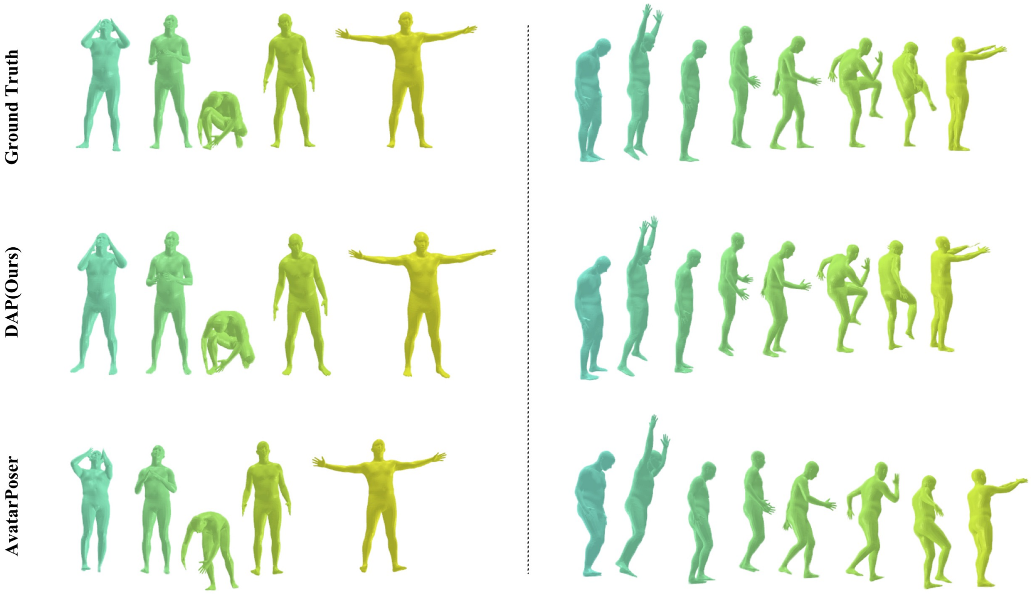

Figure 5. Visual comparisons between AvatarPoser and our method. The first row represents the ground truth of reconstruction body in a motion sequence. The second row represents the reconstruction results of our models in a motion sequence. The third row represents the results of the state-of-the-art model Avatarposer [14].

$FH^{2\to 1}$ are the cross attention features encoding the correlation between the features in the local and the global path. d_1 and d_2 are a normalization constant and f represents the function with the three features as input respectively. The cross attention features are merged into the two features Z_1 and Z_2 by a pointwise MLP layer fp respectively, which is calculated as the following:

$$FH_1^{Fusion} = fp(Z_1 + FH_1^{1\to 2}), \tag{10}$$

$$FH_2^{Fusion} = fp(Z_2 + FH_2^{2\to 1}), \tag{11}$$

where FH_1^{Fusion} and FH_2^{Fusion} are the output of features of the local and the global path after the cross attention mixer.

3.4. Attention-Gated-MLP Decoder

Two attention-gated-mlp decoders are applied to decode the features FH_1^{Fusion} and FH_2^{Fusion} of the local and the global path. Firstly, a stack of 2 blocks are applied to decode the feature FH_1^{Fusion} of the first path, each block is defined as the following:

$$H_\alpha^i = \sigma(H_{in}^i U^1) \tag{12}$$

where σ is an activate function such as GeLU [11], U_1 defines linear projection along the channel dimension for the i-th block. H_{in}^i and H_α^i represent the input and output for the σ function at the i-th block. Here, H_{in}^1 is set to FH_1^{Fusion} for the first block.

$$H_\beta^i = F_S(H_\alpha^i) \tag{13}$$

where F_S represents the function of the linear gating, the mapping is calculated as the following:

$$f_{w,b}(H_\alpha^i) = W^i H_\alpha^i + b^i \tag{14}$$

$$F_S(H_\alpha^i) = H_\alpha^i \odot f_{w,b}(H_\alpha^i) \tag{15}$$

For i-th block, W_i is a parameter matrix, b_i refers token-specific biases, $\odot$ denotes the element-wise multiplication. It's critical to initialize W_i as near-zero values and b_i as ones for the training stability, which means that $f_{w,b}\left(H_\alpha^i\right) \approx 1$ and $F_S\left(H_\alpha^i\right) \approx H_\alpha^i$ at the beginning of the training. Besides, F_s is split into two independent parts, which are proven to be more effective in training and testing [18]. The split is converted as the following:

$$F_S\left(H_\alpha^i\right) = H_{\alpha 1}^i \odot f_{w,b}\left(H_{\alpha 2}^i\right) \tag{16}$$

where H_α^i is split into two parts $\left(H_{\alpha 1}^i, H_{\alpha 2}^i\right)$ along the channel dimension for the gating function and for the multiplicative bypass.

$$H_{out}^i = H_\beta^i V^1 \tag{17}$$

Then, H_{out}^i is calculated by applying V^1 defining linear projection along the channel dimension in the i-th block.

Similarly, the operation of the feature FH_2^{Fusion} of the second path takes the same network structure and calculation process.

Further, skip connection between the first block and the second block is used for each attention gated mlp decoder. The structure of each block is shown in Figure 3.

3.5. Concatenation

Additionally, the input FH_1^{Fusion} of the first path is concatenated with the input FH_2^{Fusion} of the second path, The concatenation is calculated as the following:

$$FH_1^{concat} = FH_1^{Fusion} \oplus FH_2^{Fusion} \tag{18}$$

where the $\oplus$ represents the concatenation operator.

4. Experiments

4.1. Training Process and Dataset

Our model is trained and tested on three subsets of AMASS dataset [21] (CMU, BMLrub [33] and HDM05 [24]). Each frame of the AMASS data contains the full SMPL pose parameters (159 dimensions), which include global translations and rotations of the hand joints and the body. To adapt to VR devices, the positions and rotation matrix of head and hands are extracted from the dataset as the input to predict the full SMPL body pose. For a fair comparison with AvatarPoser [14], the same training and the same testing dataset are used. In our training procedure, Adam solver is used to optimize Dual Attention Poser, in which 40 frames of each motion are fed with batch size 256 .The learning rate starts from $1e-4$ and decays by 0.5 every $2e5$ iterations. The model is trained with Pytorch on NVIDIA GeForce GTX 3090 GPU with 24GB RAM.

4.2. Evaluation

Our model, Dual Attention Poser, is compared with the state-of-the-art method (AvatarPoser [14]) and other body tracking models including CoolMoves [1], LoBSTr [38] and VAE-HMD [6]. Similar to AvatarPoser, we use MPJRE(Mean Per Joint Rotation Error), MPJPE(Mean Per Joint Position Error), and MPJVE(Mean Per Joint Velocity Error) as the evaluation metrics. Lower evaluation metrics indicate better performance of the model. The reconstruction error is represented in Table 1.

Based on the same training and testing dataset, the test results between AvatarPoser and ours model are shown in Tabel 1, which shows that our model has better performance

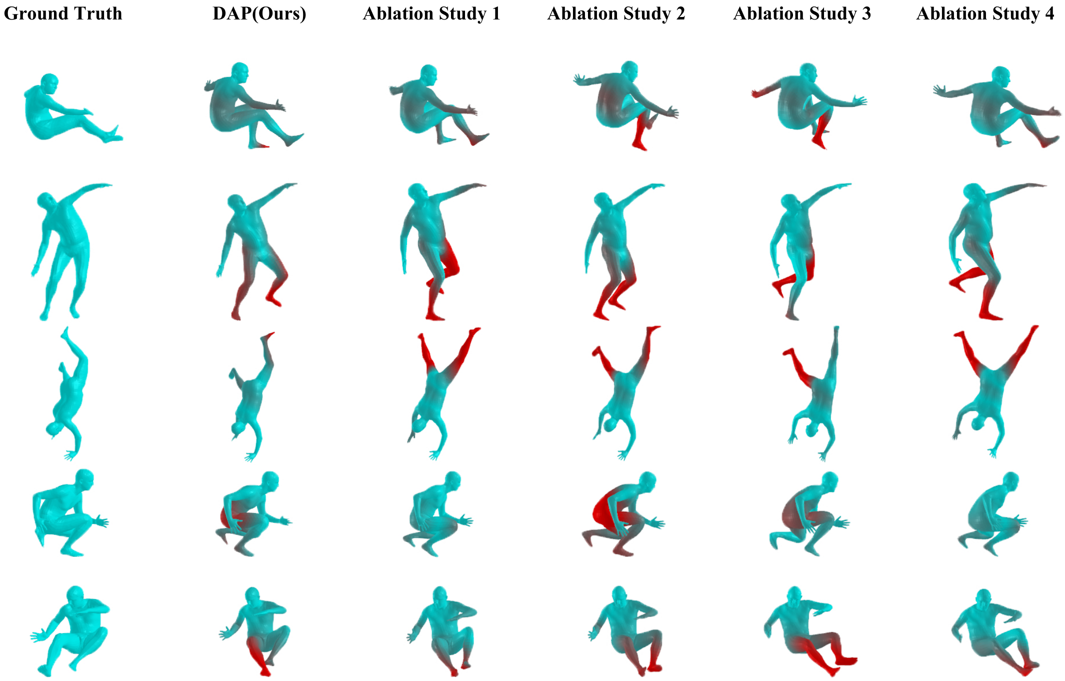

Figure 6. Visual results of Ablation Study. Ablation Study1-Ablation Study4 corresponds to "No Dual Attention Encoder", "No Cross Attention Mixer", "No Attention-Gate-MLP Decoder" and "No Concatenation".

Model	MPJRE (°)	MPJPE (cm)	MPJVE (cm/s)
CoolMoves [1]	5.20	7.83	100.54
LoBSTr [38]	10.69	9.02	44.97
VAE-HMD [6]	4.11	6.83	37.99
AvatarPoser [14]	3.21	4.18	29.40
DAP(Ours)	**2.69**	**3.68**	**24.03**

Table 1. Comparision results with the state-of-the-art methods on the AMASS dataset.

on three metrics, with MPJRE, MPJPE, and MPJVE reducing by 16.1%, 11.9% and 18.2% respectively.

To better illustrate the performance of our model, visualization experiments are conducted and the results are shown in Figure 4, in which the extent of the error is indicated by the intensity in red. The results represents that both methods have good performance in the standing and the slow walking tasks. However, for complex movements, such as running, squatting, and dancing etc, our model shows better accuracy on the prediction of body pose.

In addition, two different motion sequences are selected for the further comparisons. As shown in Figure 5, at the beginning of each sequence, both models are able to predict the body motion. With the increasing complexity of the action, our model shows the higher accuracy.

Besides, to further evaluate the generalization ability of our model, a 3-fold cross-dataset evaluation method [14] is performed among different models. The experiment result of different models tested on sub-dataset CMU, BMLrub, and HDM05 of the AMASS dataset [21] is shown in Table 2, from which the error is reduced by up to 9.6% in comparison with AvatarPoser [14].

4.3. Ablation Studies

To investigate the effectiveness of each submodule of our Dual Attention Poser, an ablation study is performed to evaluate the need for each submodule. The experiments are conducted on the same test set as Tabel 1. The results are shown in Table 3 and visualized in Figure 6, in which MPJRE, MPJPE and MPJVE are used as the evaluation metrics.

4.3.1 No Dual Attention Encoder

We remove the dual attention encoder and use a single transformer encoder to encode the input serials. According to the

Dataset	Model	MPJRE ($^\circ$)	MPJPE (cm)	MPJVE (cm/s)
CMU	CoolMoves [1]	9.20	18.77	139.17
	LoBSTr [38]	12.51	12.96	49.94
	VAE-HMD [6]	6.53	13.04	51.69
	AvatarPoser [14]	5.93	8.37	35.76
	DAP(Ours)	**5.46**	**8.15**	**32.32**
BMLrub	CoolMoves [1]	7.93	13.30	134.77
	LoBSTr [38]	10.79	11.00	60.74
	VAE-HMD [6]	5.34	9.69	51.80
	AvatarPoser [14]	4.92	7.04	43.70
	DAP(Ours)	**4.75**	**6.81**	**42.78**
HDM05	CoolMoves [1]	9.47	17.90	140.61
	LoBSTr [38]	13.17	11.94	48.26
	VAE-HMD [6]	6.45	10.21	40.07
	AvatarPoser [14]	6.39	8.05	30.85
	DAP(Ours)	**6.18**	**7.84**	**29.17**

Table 2. Results of cross-dataset evaluation in comparison with different methods.

Model	MPJRE ($^\circ$)	MPJPE (cm)	MPJVE (cm/s)
No Dual Attention	2.85	3.87	25.35
No Cross Attention	2.89	3.97	25.17
No Attention-Gate-MLP	2.83	3.92	25.41
No Concatenation	2.80	3.99	26.51
DAP(Ours)	**2.69**	**3.68**	**24.03**

Table 3. Results of the ablation studies.

results of the ablation study, MPJRE and MPJPE increased by 5.9% and 5.1%, which shows the use of dual attention encoder might help the network capture the root orientation of the body and the information of the other body pose respectively. It also enables the next Cross Attention Mixer to better integrate features of the two paths.

4.3.2 No Cross Attention Mixer

Removing the Cross Attention Mixer increases the MPJRE and MPJPE by 7.4% and 7.8% respectively, which makes our network becomes a simple encoder-decoder structure and the representation between the two paths cannot be integrated. With the removal of the Mixer, the mutual information of global and local paths cannot be effectively fused. As shown in the fourth column of Figure 6, our model cannot predict some complex body poses accurately with the Cross Attention Mixer removed.

4.3.3 No Attention-Gated-MLP Decoder

The Attention-Gated-MLP Decoder is removed and two MLP is used to decode the features of the two paths respectively, in which MPJRE and MPJPE increased by 5.2% and 6.5%. Although the increase in metrics is not significant, the visualized results in Figure 6 indicate that the Attention-Gated-MLP can help the model reduce dramatic pose errors.

4.3.4 No Concatenation

Before decoding the joint pose, we concatenate the features of the two paths in our model. The MPJRE and MPJPE are increased by 4.0% and 8.4% respectively. The error increases and the results are visualized in Figure 6.

5. Conclusion

The dual attention poser, a novel model, is proposed to estimate realistic human poses from motion signals of the mixed reality headset and the user's hands or hand-held controllers. In the network architecture, we build a dual path encoder-decoder framework to predict the 6Dof pose of landmarks of the body, According to the comparison results with the state-of-the-art models, the model proposed achieves more robust estimations results and the higher accuracy of the body reconstruction. The future work will focus on the reconstruction of the body and hand pose in one end-to-end network architecture, the exploration of efficient network architecture to reduce the latency of the current body reconstruction models and etc.

References

[1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021. 3, 6, 7, 8

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[5] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021. 3

[6] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 6, 7, 8

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[9] Andrew Goldenberg, Beno Benhabib, and Robert Fenton. A complete generalized solution to the inverse kinematics of robots. *IEEE Journal on Robotics and Automation*, 1(1):14–20, 1985. 2

[10] Keith Grochow, Steven L Martin, Aaron Hertzmann, and Zoran Popović. Style-based inverse kinematics. In *ACM SIGGRAPH 2004 Papers*, pages 522–531. 2004. 2

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[12] Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. Graph-mlp: node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*, 2021. 3

[13] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 1, 3

[14] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. *arXiv preprint arXiv:2207.13784*, 2022. 2, 3, 4, 5, 6, 7, 8

[15] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 3

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[17] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 3

[18] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021. 6

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6, 7

[22] Filip Marić, Matthew Giamou, Adam W Hall, Soroush Khoubyarian, Ivan Petrović, and Jonathan Kelly. Riemannian optimization for distance-geometric inverse kinematics. *IEEE Transactions on Robotics*, 38(3):1703–1722, 2021. 2

[23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 3

[24] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05. 2007. 6

[25] Mathias Parger, Joerg H Mueller, Dieter Schmalstieg, and Markus Steinberger. Human upper-body inverse kinematics for increased embodiment in consumer-grade virtual reality. In *Proceedings of the 24th ACM symposium on virtual reality software and technology*, pages 1–10, 2018. 2

[26] Joey K Parker, Ahmad R Khoogar, and David E Goldberg. Inverse kinematics of redundant robots using genetic algorithms. In *1989 IEEE International Conference on Robotics and Automation*, pages 271–272. IEEE Computer Society, 1989. 2

[27] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 3

[28] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 3

[29] Philipp Ruppel, Norman Hendrich, Sebastian Starke, and Jianwei Zhang. Cost functions to specify full-body motion and multi-goal manipulation tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3152–3159. IEEE, 2018. 2

[30] Robert W Sumner, Matthias Zwicker, Craig Gotsman, and Jovan Popović. Mesh-based inverse kinematics. *ACM transactions on graphics (TOG)*, 24(3):488–495, 2005. 2

[31] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021. 3

[32] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 3

[33] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002. 6

[34] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. 3

[35] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, pages 349–360. Wiley Online Library, 2017. 3

[36] L-CT Wang and Chih-Cheng Chen. A combined optimization method for solving the inverse kinematics problems of mechanical manipulators. *IEEE Transactions on Robotics and Automation*, 7(4):489–499, 1991. 2

[37] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 3

[38] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, volume 40, pages 265–275. Wiley Online Library, 2021. 1, 3, 6, 7, 8

[39] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13167–13178, 2022. 3

[40] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[41] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019. 2

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 3

[43] Jianmin Zhao and Norman I Badler. Inverse kinematics positioning using nonlinear programming for highly articulated figures. *ACM Transactions on Graphics (TOG)*, 13(4):313–336, 1994. 2

[44] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3, 4

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3