MONOCULAR 3D HUMAN POSE ESTIMATION VIA EULER ANGLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular 3D human pose estimation is a key problem in computer vision. Existing joint position-based methods often suffer from the issues of accurate bone length prediction and rotation ambiguities when joints are collinear. Joint rotation-based methods can avoid the bone length issue but encounter discontinuities when predicting body's self-rotation angles, limiting their applicability. In this work, we theoretically analyze the root cause of the discontinuity and propose a conditional Euler angle-based estimation method. Our approach projects the continuous body self-rotation angle in a high-dimensional space into a two-dimensional space and divides the angle into discrete angle intervals. A classification network learns the prior information about the body's orientations in these discrete angle intervals. Then, the orientation conditions are used as inputs to improve the prediction of the Euler angle. Experiments across diverse models, including CNNs, GCNs, and Transformers, demonstrate that our method produces continuous self-rotation prediction. It effectively resolves the discontinuity problem of Euler angles and forms a plug-and-play module for efficient model transfer.

1 INTRODUCTION

Monocular 3D human pose estimation (HPE) is a prominent research direction in computer vision with wide applications in virtual reality, augmented reality, avatar generation, and metaverse technologies(Yang et al., 2022; Anvari et al., 2023; Lu et al., 2024; Li et al., 2025). By leveraging RGB or RGB-D images and video sequences, monocular 3D human pose estimation can reconstruct the skeletal joint positions and motion in 3D space, which serves as a crucial bridge between the physical and digital worlds.

Current research can be broadly categorized into joint position-based and joint rotation-based methods. Joint position-based methods estimate the 3D coordinates of human joints in the world coordinate system to reconstruct the skeleton and motion (Zheng et al., 2023; 2021; Cai et al., 2019; Cheng et al., 2020). However, joint position-based methods face two main challenges: (1) the skeletal bone lengths often appear inconsistent, making it difficult to accurately recover joint locations. To alleviate this issue, several studies constrain bone lengths within a plausible range to improve accuracy (Chen et al., 2022; Kang et al., 2023). (2) These methods also struggle to infer self-rotation when joints are collinear. Fisch & Clark (2021) proposed virtual markers to approximate rotations along the bone axes, however, their approach remains limited by image resolution and the lack of texture information.

Joint rotation-based methods can avoid the bone length inconsistency issue (Jiang et al., 2022). However, they suffer from discontinuities caused in rotation angles by wrap-around (Pavllo et al., 2018; Zhou et al., 2019; Pepe et al., 2024). For example, after a full rotation, the angle abruptly resets to zero and leads to discontinuous predictions of self-rotation. Such wrap-around signals are difficult for neural networks to model, restricting the applicability of rotation-based methods. Common rotation representations include Euler angles (Diebel, 2006), quaternions (Pavllo et al., 2018), 6D rotation (Zhou et al., 2019), and axis-angle (Loper et al., 2015) formats. Among them, Euler angles are more intuitive and parameter-efficient compared to other forms.

In this paper, we focus on Euler angle-based pose estimation and analyze the underlying cause of discontinuities in self-rotation prediction. While the body's self-rotation is continuous in 3D space, its representation with Euler angles exhibits discontinuous jumps in the prediction. To address

this problem, we propose a conditional Euler angle-based method, in which the rotation space is projected onto a 2D plane and the angle is divided into discrete angle intervals. Using this prior interval information, we can predict continuous self-rotation. Furthermore, we design it as a plug-and-play module that can be integrated into current 3D pose estimation frameworks, including Pavllo et al. (2019); Zheng et al. (2021); Zhao et al. (2023; 2019); Liu et al. (2020b;a).

Our contributions are summarized as follows:

- We provide a theoretical analysis of the wrap-around problem in Euler angle-based representations and model it using a 3D helix. Through spatial projection, we transform continuous 3D rotation into a learnable 2D representation.
- We introduce a conditional Euler angle learning method that incorporates prior orientation information. By combining orientation with 2D joint inputs, our method enables continuous prediction and supports plug-and-play integration for existing models.
- We demonstrate efficient transfer from position-based models to rotation-based ones. Our method allows the current models to transfer from joint position to joint rotation representations, enhancing the completeness of 3D pose estimation.

The remainder of this paper is organized as follows: Section 2 reviews related works on joint position-based and joint rotation-based methods. Section 3 presents the theoretical foundation of our method and introduces conditional Euler angles for predicting continuous rotation. Section 4 presents experimental setup and results, followed by a discussion that evaluates the proposed method across a range of pose estimation models.

2 Related Work

Joint position-based estimation. (1) CNN-based methods. Pavllo et al. (2019) extended the temporal receptive field through multi-layer dilated convolutions with residual connections, thereby enhancing temporal correlations in pose sequences. Chen et al. (2021) further decoupled the task into bone length and joint direction learning to improve consistency in bone length prediction. Chen et al. (2022) leveraged bone length invariance constraints to refine 3D pose regression. To resolve rotation ambiguities in joint position representations, Fisch & Clark (2021) introduced virtual markers to model joint roll rotations. (2) GCN-based methods. Zhao et al. (2019) incorporated prior semantic information among joints to improve spatial feature modeling. Yu et al. (2023) captured global correlations via adaptive graph convolutions and refined local features through independent connection layers. Ci et al. (2019) designed a Local Connection Network (LCN) to strengthen the modeling of local spatial dependencies. (3) Transformer-based methods. Zheng et al. (2021) demonstrate the effectiveness of a pure Transformer architecture for 3D human pose estimation. Li et al. (2022) learned spatio-temporal information via multi-hypothesis generation and feature fusion. Zhang et al. (2022) achieved multi-level spatio-temporal separation and fusion by alternately stacking spatial and temporal Transformer blocks. Kang et al. (2023) proposed a dual-chain design (local-to-global and global-to-local) to fully capture complex multi-level dependencies among human joints. Shuai et al. (2023) adaptively fused multi-view and temporal features to handle varying views and video lengths without camera calibration. In summary, research on joint position-based estimation primarily focuses on temporal modeling, spatial structure constraints, and capturing both global and local dependencies. Different networks architectures (CNNs, GCNs, Transformers) have continuously improved the representation of spatio-temporal dependencies among joints. Some works attempt to alleviate issues, such as inconsistent bone lengths and missing self-rotation, but these challenges remain largely unsolved.

Joint rotation-based estimation. Akhter & Black (2015) investigated the rotation limits of human joints under specific poses and introduced limit-value constraints. Building on this, Yang et al. (2023) incorporated human pose priors via learnable rotation tokens to constrain estimated angles within plausible ranges. To mitigate error accumulation along kinematic chains, Pavllo et al. (2018) employed recurrent neural networks to estimate joint rotations and introduced a differentiable forward kinematics loss to mitigate accumulated errors. However, joint rotation representations suffer from discontinuities. Several strategies have been proposed to solve the problem. Burgermeister & Curio (2022) adopted spherical coordinates, with the polar angle represents pitch and the azimuth angle represents horizontal orientation. Li et al. (2021) decomposed rotations into twist and swing

components, estimated respectively via numerical computation and neural prediction. Zhou et al. (2019) proposed the 6D rotation representation and Banik et al. (2024) further utilized 2D rotation information from 2D keypoints to assist 3D rotation estimation. In summary, discontinuities in rotation-based representations are typically addressed in two ways: (1) Reducing rotational degrees of freedom using adopting 2D angle representations, which avoids discontinuities but sacrifices one degree of freedom, limiting the ability to represent complex rotations; (2) Employing continuous representations, which preserve full rotational information but increase parameter complexity and may introduce errors.

3 METHOD

3.1 THESIS ANALYSIS

To regress the sequence of human pose, we aim to learn a continues function for the body rotation:

$$\theta = f_w(x) \tag{1}$$

Here, θ lies on a helix, with body rotation defined over [0, $+\infty$). The ground-truth value typically falls within $[\theta, 2n * 180]$, where n denotes the number of full rotations. However, from a single image, $f_w(x)$ can only predict from 0° to 360°, since the total number of full rotations cannot be inferred without prior information.

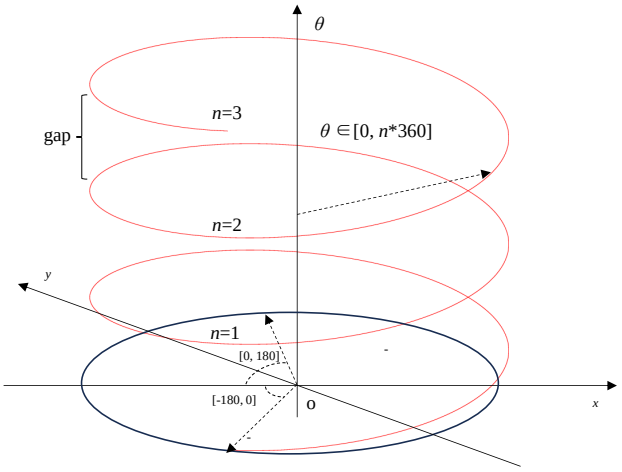


Figure 1: The body rotation of the human pose follows a helical trajectory.

As shown in Figure 1, there is a gap between each round in a helix. The network predicts the continuous rotation value, which can causes the issues of wrap-around at the gap. By projecting the helix into 2D Cartesian coordinates, the regression interval from $[0, 2n * 180)$ is mapped to [0, 360). After shifting 180°, the range can be divided into two intervals: [-180, 0), where y is negative, and [0, 180), where y is positive. To generalize the formula, Eq. (1) can be rewritten as below:

$$\theta = f_w(x, y) \tag{2}$$

This modification allows the network to learn continuous rotations within [0, 360) by incorporating an additional feature that indicates the sign of y.

3.2 CONDITIONAL POSE ESTIMATION

The overall framework of conditional Euler angle-based pose estimation is illustrated in Figure 2. Our method consists of three components: (1) Condition inference: Given the input 2D keypoints, the model infers the horizontal rotation interval of the body's orientation and generates a corresponding condition label. (2) Conditional Euler angle-based pose estimation: The inferred condition is

fused with the original 2D keypoints to construct an augmented keypoints representation. The augmented keypoints are then fed into existing pose estimation networks in a plug-and-play manner, enabling Euler angle-based human pose regression. (3) Neural Networks: The method is compatible with all existing network architectures for 3D human pose estimation, including those models based on CNN, GCN, and Transformer.

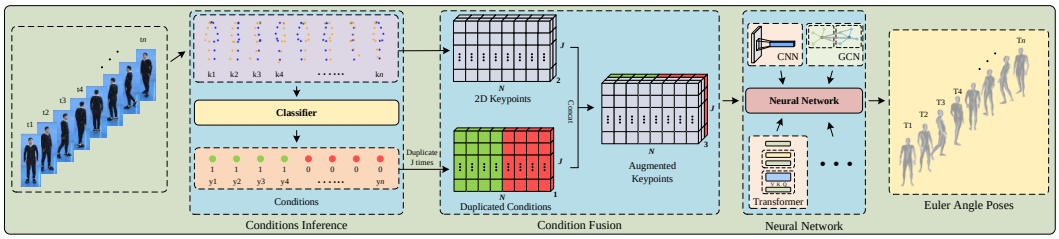


Figure 2: The overall framework of conditional Euler angle-based human pose estimation.

3.2.1 Learning Condition

In rotation-based methods, the overall body rotation is determined by the horizontal rotation of the root joint. The horizontal rotation angle can increase continuously and span multiple cycles. Therefore, it is difficult to estimate the exact number of full rotations using only a limited set of frames.

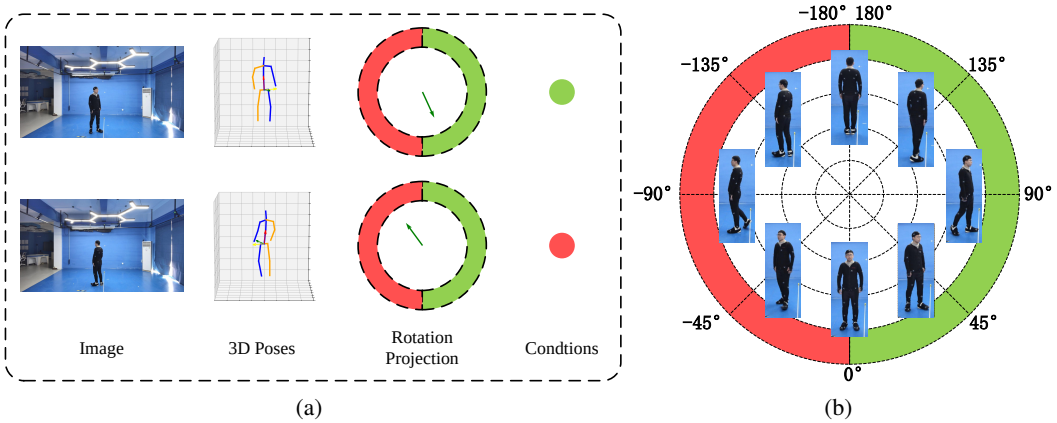


Figure 3: (a): Rotation Conditions. The horizontal Euler angle is divided into two intervals based on the body's orientation. (b) Angular intervals corresponding to the conditional Euler angles. The green interval spans [0, 180], while the red interval covers [-180, 0].

As illustrated in Figure 3(b), we implement a conditional Euler angle learning method by projecting the root joint's angle onto the 2D plane, allowing the angle can be divided into two intervals according to the body's orientation. In Figure 3(b), we present the condition results of different human body orientations. For representing the self-rotation of each pose in the camera view, the projected angle in the world coordinate system is denoted as $\phi \in [-180, 180]$. Based on the interval of ϕ, a conditional label y is assigned as follows:

$$y = \begin{cases} 1, & \phi \geq 0 \\ 0, & \phi < 0 \end{cases} \tag{3}$$

When the body faces the right side of the camera, the projected angle lies in [0, 180], the conditional label y is set to 1. When the body faces the left side, the projected angle lies in $[-180, 0]$, and y is set to 0. This conditional definition provides an explicit constraint that helps mitigate discontinuities in rotation prediction. A classifier is employed to infer the body's orientation condition. Given the

input 2D keypoints (u,v), the classifier predicts the current condition y, as illustrated in Eq. (4) and Figure 4, where the 2D keypoints are processed to get the corresponding orientation condition.

$$y = \text{classifier}(u, v) \tag{4}$$

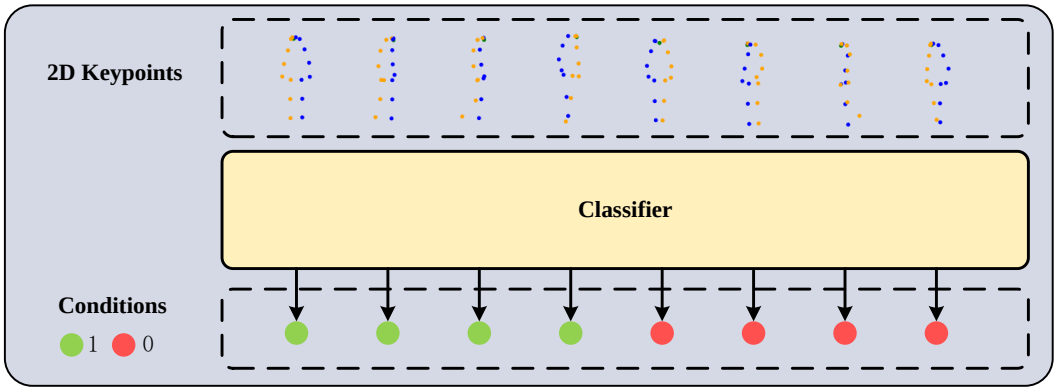


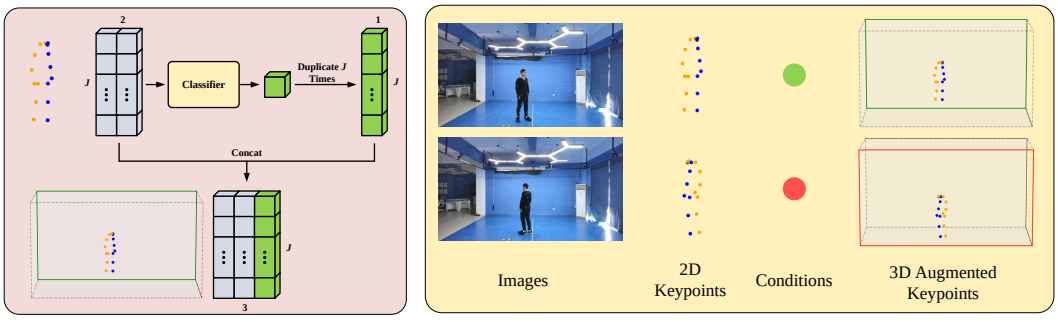
Figure 4: The classifier infers the body's orientation condition from the 2D keypoints.

3.2.2 Condition Fusion

As described in Eq. (2), the condition scalar c is fused with the 2D keypoints. For a sequence of N frames, the condition scalar $c \in \mathbb{R}^{N\times 1}$ is replicated along the joint dimension to obtain $c_J \in \mathbb{R}^{N\times J\times 1}$. Then concatenation operation combines c_J with the original 2D keypoints P along the feature dimension to form the extended input $P_{\text{ex}} \in \mathbb{R}^{N\times J\times 3}$, as shown in Eq. (5).

$$P_{\text{ex}} = \text{concat}(P, c_J) \in \mathbb{R}^{N\times J\times 3} \tag{5}$$

For each joint in each frame, the condition value is appended to the corresponding 2D keypoint coordinates, forming an extended feature vector of dimension 3 per joint. In Figure 5(b), It can be seen that the 2D keypoints are projected onto the plane of $y = 1$ or the plane of $y = 0$ according to the inferred conditions. Thereby, by embedding discontinuous signals of Euler angles at the input, this method substantially enhances the model's capability to learn and predict Euler angle outputs continuously.



(a) (b)

Figure 5: Schematic diagram of conditional Euler angles. (a) 2D keypoints and conditional information are concatenated to form dimensionally augmented 3D keypoints. (b) The condition label maps the 2D keypoints onto different planes. The 2D keypoints lie in either the $y = 0$ plane (green rectangular box) or the $y = 1$ plane (red rectangular box).

3.2.3 Loss Function

For rotation-based pose estimation, we adopt the Mean Per Joint Angular Sepration Error (MPJAE) as the loss function, as defined in Eq. (6).

$$\text{MPJASE} = \frac{1}{N \times J} \sum_{i=1}^{N} \sum_{j=1}^{J} \|e_{i,j} - \hat{e}_{i,j}\|_1 \tag{6}$$

Eq. (6) quantifies the average L_1 distance between the predicted Euler angle $\hat{e}_{i,j}$ and the ground-truth Euler angle $e_{i,j}$ across all joints j and frames i, each angle contains (α, β, γ) three scales, where N denotes the number of frames and J denotes the number of joints. This metric thus measures the rotational discrepancy between prediction and ground truth.

4 Experimental Results and Discussion

To validate the effectiveness of the proposed method, we examined its performance on our dataset. Since existing public datasets do not provide Euler angle annotations, we collected our own dataset, with details presented in the appendix A. The experiments consist of two parts: (1) evaluating the performance of different classifiers for condition inference; (2) embedding angle conditions into existing 3D pose estimation models to assess their ability to handle rotational discontinuities.

4.1 Classifier Study

The condition inference is performed based on 2D keypoints obtained using the Detectron2 2D pose estimation framework. To incorporate temporal information, the input is represented as a sequence of N frames, each containing J joints, resulting in $P \in \mathbb{R}^{N \times J \times 2}$. For optimization, cross-entropy loss is employed to guide the network in discriminating condition labels. Ground-truth body orientations are manually annotated, with positive values assigned to instances facing left and negative values to those facing right.

We compare ResNet and a softmax classifier for condition inference. As shown in Table 1, both classifiers achieve over 98% accuracy on the test set and demonstrate a stable discrimination for the body orientations. While their accuracies are comparable, the Softmax classifier offers faster inference speed.

Table 1: Performance of two classifiers on conditional inference tasks

Classifier	**Accuracy**	**Speed(frame/s)**
ResNet18	98.3%	96212
Softmax	98.1%	608364

As shown in Table 1, the ResNe18 network achieves 98.3% accuracy at a speed of 96,212 frames/s, whereas the softmax classifier achieves 98.1% accuracy at a speed of 608,364 frames/s. However, the two classifiers exhibit different patterns in misclassification.

Figure 6 illustrates examples where the ResNet18 network and the Softmax classifier predict correctly, and both fail. In Figure 6(a), frames correctly classified by ResNet18 predominantly correspond to standing actions, whereas frames correctly classified by Softmax classifier mostly correspond to sitting or squatting actions as shown in Figure 6(b). In Figure 6(c) the misclassified by both classifiers typically involve heavily occluded actions. These observations indicate that the classifiers differ across action types and both struggle under conditions of extreme occlusion. Here, the Resnet18 network was used for further study.

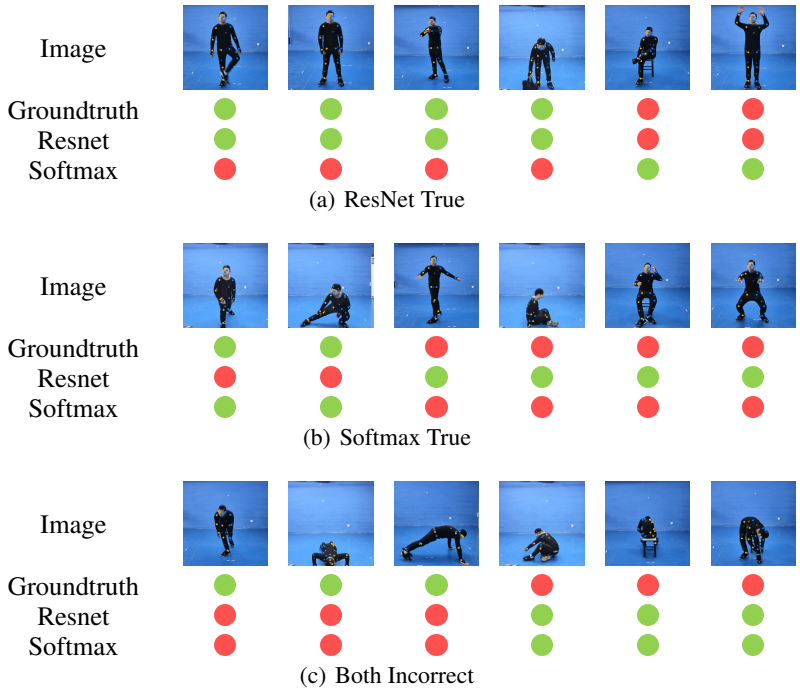


Figure 6: Comparison of classifier predictions. (a) Frames correctly classified only by the ResNet18 network. (b) Frames correctly classified only by the softmax classifier. (c) Frames misclassified by both classifiers. Colored circles indicate the prediction results, green means true, while red means false.

4.2 Experimental Validation on Multiple Baselines

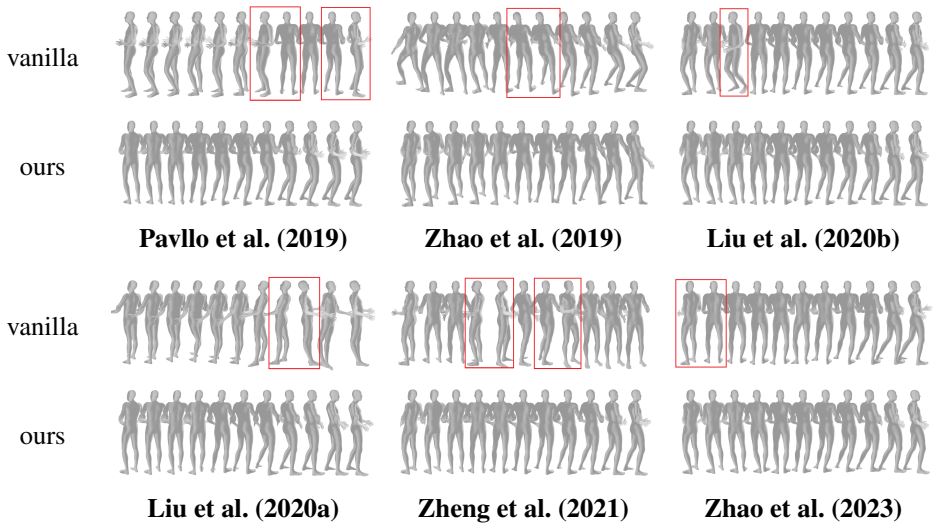
To further verify the effectiveness of our method, we applied it to six representative 3D human pose estimation methods: Pavllo et al. (2019), Zhao et al. (2019), Liu et al. (2020b), Zheng et al. (2021),Zhao et al. (2023), and Liu et al. (2020a). All methods share identical preprocessing steps, input-output format, and training strategy, and adopt the same loss function (MPJASE) for comparability. Each model was trained on the same training set and evaluated on the same test set. Detailed settings are provided in Appendix C.

Table 2: Quantitative comparison of MPJASE between the vanilla (before) and ours (after) on our angle Dataset.

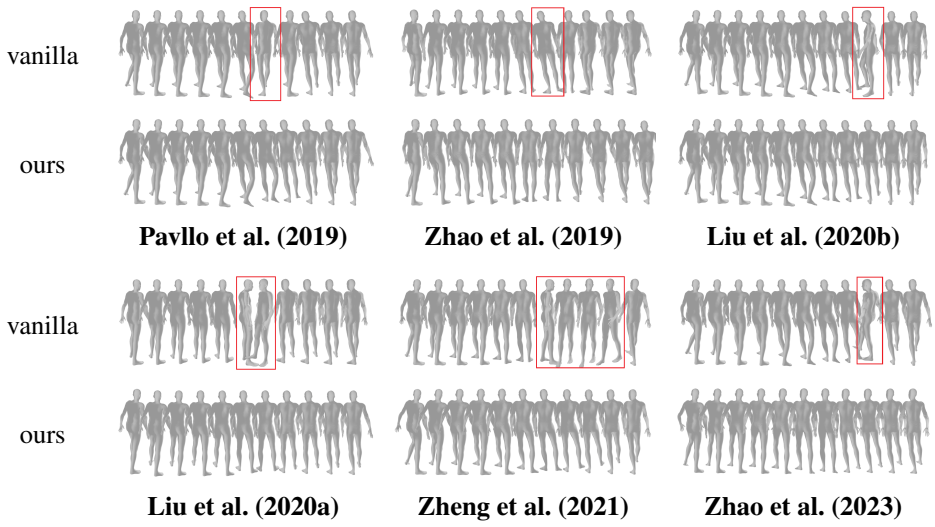
Methods	**Walk**	**Sit**	**Run**	**Jump**	**Squat**	**Torso**	**Arm**	**Leg**	**Avg**
Pavllo et al. (2019)	5.73/**5.16**	8.00/**7.32**	5.99/**5.50**	6.50/**5.44**	5.46/**5.28**	6.47/**6.31**	6.00/**5.81**	7.26/**6.61**	6.40/**5.95**
Zhao et al. (2019)	10.32/**8.59**	11.25/**10.77**	10.84/**8.69**	11.13/**9.20**	10.37/**9.20**	11.34/**9.90**	10.27/**8.54**	10.96/**9.78**	10.81/**9.33**
Zheng et al. (2021)	6.90/**5.15**	8.49/**7.41**	7.77/**5.32**	7.05/**5.18**	5.81/**5.60**	8.07/**6.58**	6.31/**6.13**	8.61/**7.10**	7.38/**6.06**
Zhao et al. (2023)	9.19/**5.90**	11.44/**8.38**	9.72/**5.62**	8.23/**6.35**	6.77/**6.75**	9.01/**7.47**	7.52/**6.96**	9.65/**7.79**	8.90/**6.90**
Liu et al. (2020b)	5.88/**5.65**	7.87/**7.80**	5.61/**5.57**	6.13/**6.01**	6.12/**6.10**	7.13/**7.06**	6.36/**6.10**	7.58/**7.35**	6.6/**6.40**
Liu et al. (2020a)	7.05/**5.58**	8.02/**7.77**	6.62/**5.22**	8.42/**6.10**	6.19/**6.11**	8.22/**7.24**	6.75/**6.33**	9.34/**7.31**	7.60/**6.50**

Table 2 presents the MPJASE results across different action categories before and after incorporating our method. Pavllo et al. (2019) achieves an average reduction of 0.5°, with notable improvements on *Jump* (1.06°). Zhao et al. (2019) despite having the highest overall error, also benefits significantly with an average decrease of 1.48°. Zheng et al. (2021) shows pronounced gains on dynamic actions, particularly *Walk* (1.75°) and *Run* (2.45°). Zhao et al. (2023) achieves the largest improvement of 2.04°.Liu et al. (2020b) achieves a moderate improvement (average 0.11°) while still maintaining consistent gains. Liu et al. (2020a) exhibits more evident improvements on *Walk* (1.47°) and *Jump* (2.32°). Actions such as *Walk*, *Jump*, and *Run* show the most significant reductions of 1.77°,

1.52°, and 1.51°, respectively, indicating that our method is particularly effective for multi-cycle rotations. In contrast, *Squat* and *Torso*, which primarily involve static or localized movements, exhibit relatively smaller improvements of 0.28° and 0.53°. These results demonstrate that our method is especially advantageous in handling dynamic body's rotations, without any degradation in performance.



(a) Qualitative comparison of vanilla and ours in *Run* action



(b) Qualitative comparison of vanilla and ours in *Walk* action

Figure 7: Qualitative comparison of vanilla and ours across different methods for *Run* and *Walk* actions.

We further analyze the performance of different methods at rotational discontinuities. Figure 7 shows consecutive frames of *Run* and *Walk* actions during abrupt rotation changes. In the *Run* action (Figure 7(a)), Pavllo et al. (2019) and Liu et al. (2020a) exhibit sharp rotational jumps without smooth transitions, while Zheng et al. (2021) shows over- or under-rotation across several frames. Liu et al. (2020b) and Zhao et al. (2023) generally produce smoother transitions but still have frame-level errors, whereas Zhao et al. (2019) causes unnatural body tilting. In the *Walk* action (Figure 7(b)), Pavllo et al. (2019); Liu et al. (2020b); Zheng et al. (2021); Zhao et al. (2023); Liu et al. (2020a) display directional errors in some frames, with Zhao et al. (2019) again exhibiting tilting. These observations highlight the limitations of Euler-angle representations, which often yield dis-

continuous predictions at interval boundaries. By incorporating conditional estimation, both actions achieve smooth and coherent transitions across all six methods, demonstrating the effectiveness and generality of our approach.

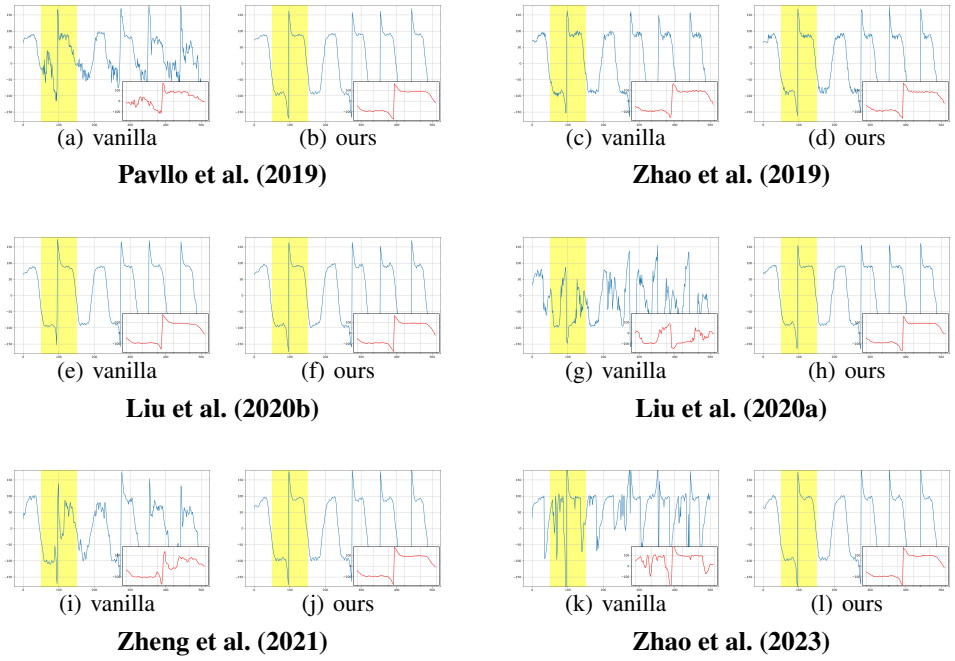


Figure 8: Comparison of horizontal rotation angles. The yellow region indicates the angle range from $-180°$ to $180°$. Pavllo et al. (2019); Zheng et al. (2021); Zhao et al. (2023); Liu et al. (2020a) fail to capture the step changes, Zhao et al. (2019); Liu et al. (2020b) exhibit oscillations. With our method, all models produce continuous rotations.

To further validate the effectiveness of our method, Figure 8 illustrates the curves of horizontal rotation angle curves over frames for different approaches. The results reveal a typical periodic wrap-around phenomenon. Without conditional estimation, the predicted curves fail to correctly follow the ground truth at step changes. Specifically, Pavllo et al. (2019); Zheng et al. (2021); Zhao et al. (2023); Liu et al. (2020a) exhibit strong fluctuations around the discontinuities, indicating difficulty in modeling rotational step changes. Zhao et al. (2019); Liu et al. (2020b) produce curves closer to the ground truth, but their step transitions occur a few frames earlier, resulting in inaccurate pose predictions. After applying our method, the predictions accurately reconstruct the step transitions. These results confirm that the proposed conditional method effectively address the challenge of managing discontinuous in Euler angle-based pose estimation.

5 Conclusion

This paper presented a conditional Euler angle-based method for 3D human pose estimation, addressing the discontinuity problem of wrap-around in body self-rotation angle prediction. By leveraging human orientation priors, the method projects the self-rotation space of the human body onto a conditional plane, which can be seamlessly integrates into existing 3D human pose estimation frameworks in a plug-and-play manner. Experiments demonstrate significant performance gains across current CNN-, GCN-, and Transformer-based models. Our study demonstrates that explicitly modeling conditional information can effectively enhance the prediction continuity in rotation representation and the generalization of human pose estimation. Future work will extend the method to parametric human body models and explore novel neural network structures to improve the prediction accuracy, Overall, incorporating orientation priors enables Euler rotation angles to be learned reliably, effectively resolving the wrap-around problem.

REFERENCES

Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Taravat Anvari, Kyoungju Park, and Ganghyun Kim. Upper body pose estimation using deep learning for a virtual reality avatar. *Applied Sciences*, 2023.

Soubarna Banik, Edvard Avagyan, Sayantan Auddy, Alejandro Mendoza Gracia, and Alois Knoll. Posegraphnet++: Enriching 3d human pose with orientation estimation. *arXiv preprint arXiv:2308.11440*, 2024.

Dennis Burgermeister and Cristóbal Curio. Pedrecnet: Multi-task deep neural network for full 3d human pose and orientation estimation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022.

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Shu Chen, Yaxin Xu, Zhengdong Pu, Jianquan Ouyang, and Beiji Zou. Skeletonpose: Exploiting human skeleton constraint for 3d human pose estimation. *Knowledge-Based Systems*, 2022.

Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019.

James Diebel. Representing attitude: Euler angles, unit quaternions, and rotation vectors. *Matrix (Stuttgart, Germany)*, 2006.

Martin Fisch and Ronald Clark. Orientation keypoints for 6d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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. In *European conference on computer vision (ECCV)*, 2022.

Hongbo Kang, Yong Wang, Mengyuan Liu, Doudou Wu, Peng Liu, and Wenming Yang. Double-chain constraints for 3d human pose estimation in images and videos. *arXiv preprint arXiv:2308.05298*, 2023.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 (CVPR)*, 2022.

Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multimodal framework for human pose comprehension, generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Junfa Liu, Juan Rojas, Zhijun Liang, Yihui Li, and Yisheng Guan. A graph attention spatio-temporal convolutional networks for 3d human pose estimation in video. *arXiv preprint arXiv:2003.14179*, 2020a.

Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 2015.

Feichi Lu, Zijian Dong, Jie Song, and Otmar Hilliges. Avatarpose: Avatar-guided 3d pose estimation of close human interaction from sparse multi-view videos. In *European Conference on Computer Vision (ECCV)*, 2024.

Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference (BMVC)*, 2018.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 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 (CVPR)*, 2019.

Alberto Pepe, Joan Lasenby, and Pablo Chacón. Learning rotations. *Mathematical Methods in the Applied Sciences*, 2024.

Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jianfei Yang, Yunjiao Zhou, He Huang, Han Zou, and Lihua Xie. Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation. In *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022.

Sen Yang, Wen Heng, Gang Liu, Guozhong Luo, Wankou Yang, and Gang Yu. Capturing the motion of every joint: 3d human pose and shape estimation with independent tokens. 2023.

Bruce X. B. Yu, Zhi Li Zhang, Yongxu Liu, S. Zhong, Y. Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human. *ArXiv*, 2023.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 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 (CVPR)*, 2022.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 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 (CVPR)*, 2023.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM computing surveys*, 2023.

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 (CVPR)*, 2019.

A DATASET

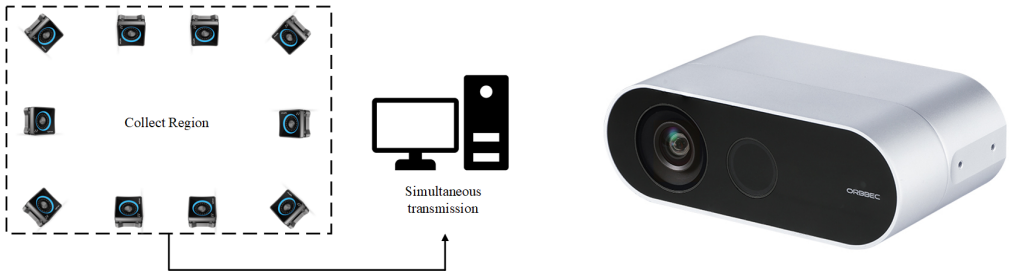
In this study, we constructed a human action dataset represented by 3D Euler angle sequences. Several participants were enrolled and performed motions after training within a calibrated capture space of $5m$×$3.2m$×$2.5m$. High-precision body motion was captured using an OptiTrack system, with synchronized video sequences were recorded with an Orbbec Femto Bolt camera. Each subject was equipped with 41 reflective markers, the trajectories of which were subsequently converted into Euler angle rotations for 17 body joints.

The dataset covers a diverse set of actions, including *Walk*, *Sit*, *Run*, *Jump*, *Squat*, *Torso*, *Arm*, and *Leg*, spanning extreme rotational angles across all joints. Human poses are represented by Euler angle rotations of 17 joints together with a 3D displacement vector of the root joint. During data acquisition, the optical axis of the camera was aligned with the $-z$ direction of the motion capture coordinate system, ensuring that horizontal body rotations are accurately projected onto the imaging plane. Examples of captured actions are illustrated in Figure 9, while the motion capture setup and equipment are shown in Figure 10. The number of frames per action is summarized in Table 3.

Table 3: Frame distributions of the training and test sets.

	Walk	**Sit**	**Run**	**Jump**	**Squat**	**Torso**	**Arm**	**Leg**	**Sports**	**Sum**
Training set	13993	17843	2168	5253	2613	11027	26031	12246	20151	117325
Test set	1229	1963	1040	988	275	959	2429	1112	22924	32219

Figure 9: Examples of actions in our datase



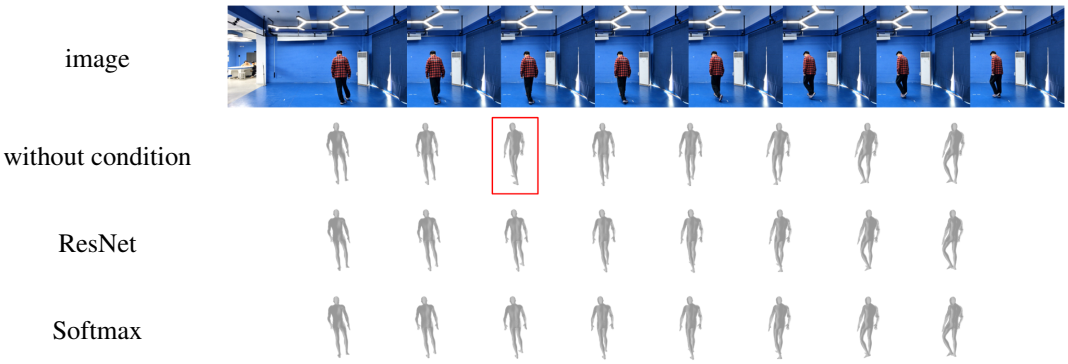
(a) OptiTrack Motion capture system

(b) Orbbec Femto Bolt Camera

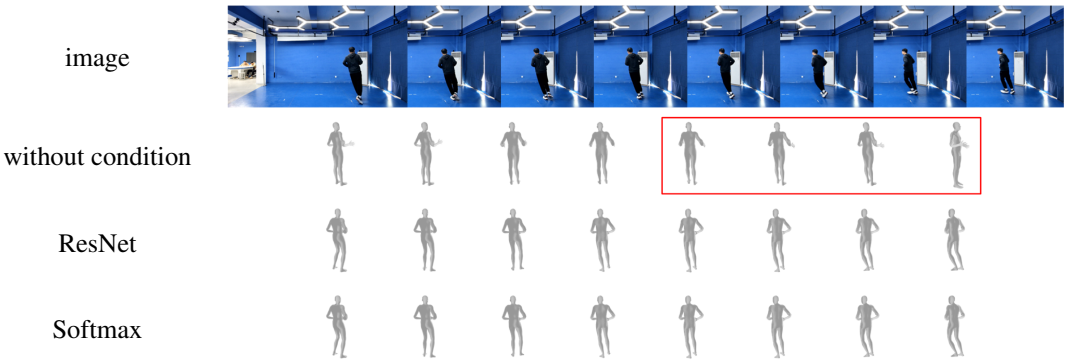
Figure 10: The motion capture system and camera for our dataset collection

B INFERENCE VIDEOS

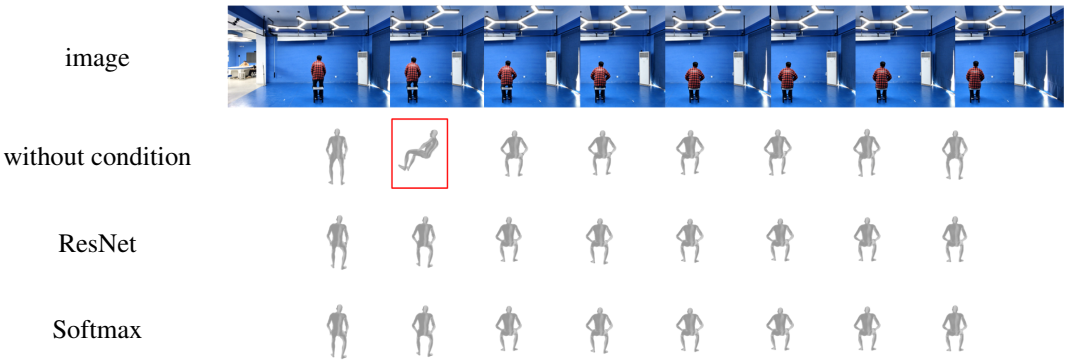
We present several inference results on out-of-dataset videos. As illustrated in Figure 11, the classifier-inferred conditional information is applied to the framework from Pavllo et al. (2019). Without the conditional information, noticeable discontinuities or reversed rotations appear in the joint rotation sequences (highlighted by red boxes). Incorporating conditional inference effectively corrects these discontinuities. These results demonstrate the ability of the classifier to directly address discontinuity issues in pose estimation.



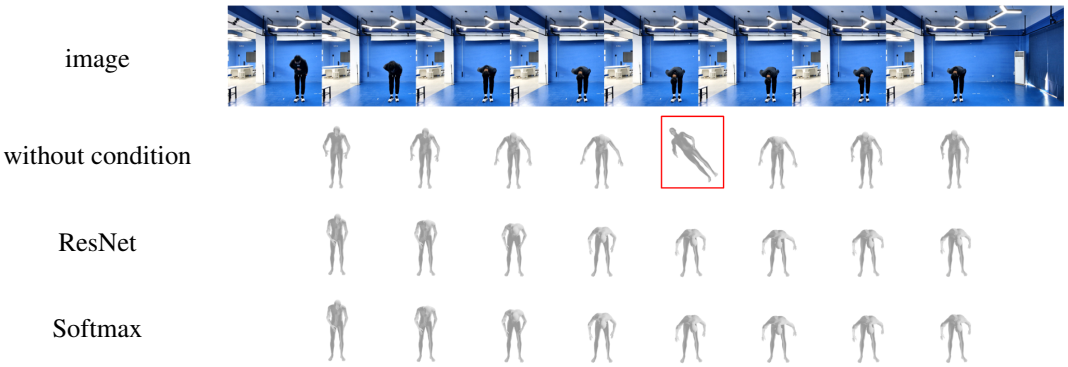
(a) *Walk* action



(b) *Run* action



(c) *Sit* action



(d) *Bow* action

Figure 11: Results based on classifier-inferred conditions.

C TRAINING DETAILS

To accommodate training with Euler angle rotation-based human pose dataset, we applied uniform preprocessing to all methods.: (1) Preprocessing operations originally designed for position learning were replaced with angle-based preprocessing to match the Euler angle representation. (2) The input channel dimensions of the networks were adjusted. The original networks were designed for 2D joint inputs; additional channels were added to incorporate the conditional Euler angle priors.(3)The loss functions were modified by replacing the original MPJPE with MPJASE for all methods, providing a more appropriate metric for Euler angle errors.

All baseline methods are implemented with their official default configurations or the authors' recommended setups, while our proposed module is integrated into each of them for evaluation. To ensure fairness, we follow the original training protocols as closely as possible and only make minimal adjustments where necessary. Specifically, Pavllo et al. (2019) is trained with five layers of dilated convolutions (kernel size 3) and a batch size of 1024, without strided convolutions, causal convolutions, or semi-supervised learning, for a total of 60 epochs. Zheng et al. (2021) is trained with $f = 81$ using its default settings until convergence, and Zhao et al. (2023) adopts $f = 27$ with similar hyperparameters and a comparable training schedule. Liu et al. (2020b) is trained for 40 epochs using its default configuration. Liu et al. (2020a) is trained with a batch size of 128 and default architectural parameters for a moderate number of epochs to ensure stable convergence.Zhao et al. (2019) is trained with the non-local module enabled for 90 epochs to capture spatiotemporal dependencies. For all methods, the initial learning rate is set to 0.001 and decayed by a factor of 0.95 per epoch, consistent with the respective default schedules.

Table 4: Training configurations for all 3D human pose estimation methods.

Method	**Core Settings**	**Batch Size**	**Epochs**	**Special Modules**
Pavllo et al. (2019)	$arc = 3, 3, 3, 3, 3$	1024	60	–
Zheng et al. (2021)	$f = 81$	default	70	–
Zhao et al. (2023)	$f = 27$	default	70	–
Liu et al. (2020b)	default	default	40	–
Liu et al. (2020a)	$arc = 3, 3, 3$	128	60	–
Zhao et al. (2019)	default	default	90	non-local module

**Note: All methods use initial learning rate lr = 0.001 with decay factor 0.95 per epoch.*