EVOPOSE: A RECURSIVE TRANSFORMER FOR 3D HUMAN POSE ESTIMATION WITH KINEMATIC STRUCTURE PRIORS

Yaqi Zhang[1,2], *Yan Lu*[3], *Bin Liu*[1,2⋆], *Zhiwei Zhao*[1,2], *Qi Chu*[1,2], *Nenghai Yu*[1,2]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]CAS Key Laboratory of Electromagnetic Space Information [3]The University of Sydney

ABSTRACT

Transformer is popular in recent 3D human pose estimation, which utilizes long-term modeling to lift 2D keypoints into the 3D space. However, current transformer-based methods do not fully exploit the prior knowledge of the human skeleton provided by the kinematic structure. In this paper, we propose a novel transformer-based model EvoPose to introduce the human body prior knowledge for 3D human pose estimation effectively. Specifically, a Structural Priors Representation (SPR) module represents human priors as structural features carrying rich body patterns, e.g. joint relationships. The structural features are interacted with 2D pose sequences and help the model to achieve more informative spatiotemporal features. Moreover, a Recursive Refinement (RR) module is applied to refine the 3D pose outputs by utilizing estimated results and further injects human priors simultaneously. Extensive experiments demonstrate the effectiveness of EvoPose which achieves a new state of the art on two most popular benchmarks, Human3.6M and MPI-INF-3DHP.

Index Terms— 3D human pose estimation, Transformer, kinematic structure, recursive refinement

1. INTRODUCTION

Monocular 3D human pose estimation (HPE) aims to estimate 3D joint positions of the human skeleton from given videos, which has rich practical application scenarios, such as motion capture [1, 2] and virtual reality [3]. An effective pipeline is separating the 3D HPE as a two-stage system consisting of 2D keypoints detection and 3D joints lifting [4]. Between them, the key barrier is the 3D joints lifting because of the ill-posed property caused by the depth leakage. Existing state-of-the-art 3D HPE methods [5, 6, 7] solve the lifting problem by modeling long-term and fine-grained spatiotemporal information with Transformer [8] and achieve advanced results. But they often merely extract features for 2D pose sequences, which misses the human priors and limits their performances.

We believe that human priors, also called kinematic structural information, can provide much more useful information to constrain the estimated human poses, which has the potential to lead to more reliable results. In particular, kinematic cues of human skeleton [9] provide body structural knowledge that indicates the category of each joint and the connectivities [9] for every joint pair, which results in more plausible 3D poses in physical. To allocate that prior knowledge more effectively in the Transformer model for 3D HPE, we propose a novel transformer lifting model EvoPose.

EvoPose includes a Structural Priors Representation (SPR) module, a SpatioTemporal Enhancement (STE) module, and a Recursive Refinement (RR) module to introduce human priors effectively. Firstly, the SPR extracts structural features from the kinematic tree to introduce human priors initially. And then in the STE module, the 2D pose sequence is combined with the structural features in a STEvo block to inject the kinematic constraint in the spatiotemporal modeling process, leading to more informative spatiotemporal sequence features. Finally, the RR module estimates 3D poses based on these sequence and structural features recursively to further incorporate the human priors. With the aforementioned processes, the human priors have been introduced to 3D HPE step-by-step. Our main contributions are as follows:

- We present a novel transformer method, EvoPose, for monocular 3D HPE using three modules to introduce kinematic structure priors effectively step-by-step.
- Our EvoPose achieves state-of-the-art performance on two datasets, surpassing existing methods by 10.3mm under MPJPE on MPI-INF-3DHP especially.

2. METHODS

The overview of our proposed EvoPose is illustrated in Fig. 1. Firstly, a Structural Priors Representation (SPR) module formulates the kinematic tree as structural features P to indicate the high-level relations in each joint pair. Then, a transformer module SpatioTemporal Enhancement (STE) guides S and P to interact with each other and extracts stronger sequence and structural features (defined as S^e and P^e) for further usage. Finally, a Recursive Refinement (RR) module estimates 3D poses by modeling on S^e and P^e recursively.

⋆Corresponding author, email: flowice@ustc.edu.cn.

This work is supported by the National Natural Science Foundation of China (Grant No. 62272430).

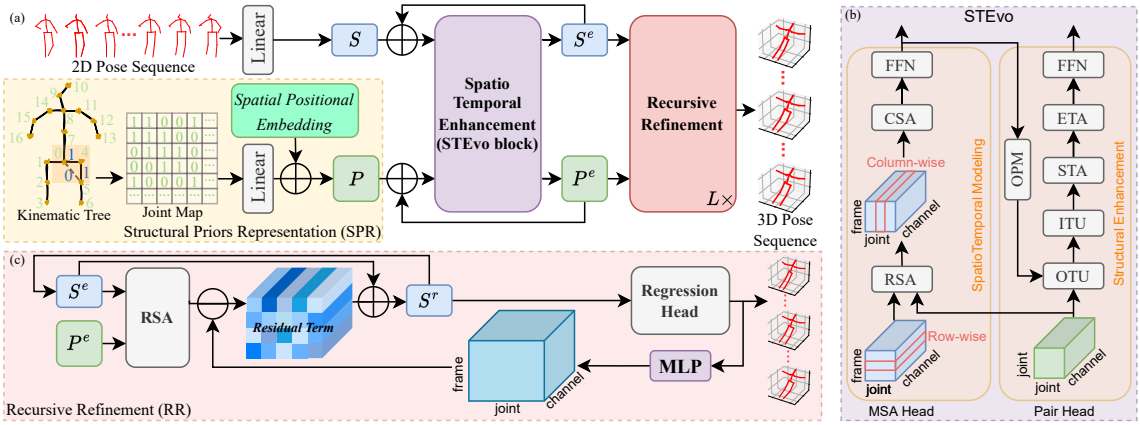

Fig. 1: **(a)** Overview of the proposed EvoPose. Structural Priors Representation (SPR) derives structural features from the kinematic tree. SpatioTemporal Enhancement (STE) models spatiotemporal relations and enhances structural features. **(b)** Overview of STEvo in STE. **(c)** Overview of the Recursive Refinement (RR) which refines the estimated result auto-regressively.

2.1. Structural Priors Representation

To introduce the human priors into our model, the SPR first builds a joint map $M \in \mathbb{R}^{J\times J\times 1}$ as Fig. 1(a) shows and J is the total number of body joints. Specifically, if joint i and j are connected in the kinematic tree, then $M_{ij} = M_{ji} = 1$, otherwise $M_{ij} = M_{ji} = 0$. This map represents joints connection in the kinematic tree, indicating the local relationships in the human body. We further use a Spatial Positional Embedding (SPE) to introduce the global kinematic cues. Because the ordering of joints follows the rules of the kinematic tree, the gap of joint number can reflect the category difference in each joint pair, providing more global relationships. We formulate these gaps into one-hot vectors and map them to SPE as a tensor with the shape of $J \times J \times d_p$, where d_p is the structural feature dimension. After that, the joint map is fed into a linear layer to expand its dimension and added with the SPE to get the structural features $P \in \mathbb{R}^{J\times J\times d_p}$.

2.2. SpatioTemporal Enhancement

When the SPR processes human body priors, the 2D pose sequence is also transferred to initial features $S \in \mathbb{R}^{N\times J\times d_s}$, where N is the frame number and d_s is the sequence feature dimension. To integrate human priors into the spatiotemporal modeling process, we propose a SpatioTemporal Evoformer (STEvo) block into the STE module to incorporate the human priors P and the sequence features S effectively.

2.2.1. *SpatioTemporal Evoformer*

The original Evoformer [10] has two input heads, a multiple sequence alignment (MSA) representation one, and a pair representation one. In the STEvo, we send the temporal features S to the MSA head and the human structural features P to the pair representation head, which is shown in Fig. 1(b). Obviously, our inputs have quite different meanings from the original Evoformer, but they share the same dimension formats. Specifically, the pose features S, sent to the MSA head, have the shape of $N \times J \times d_s$, and the structural features P, fed to the pair representation head, possess $J \times J \times d_p$ shape in our STEvo. The message interactions of S in STEvo can be interpreted as two levels: the spatial level on J joints and the temporal level across N frames. In this process, human priors provided by P can be injected into the pose features S effectively and P are also interacted with S, leading to more expressive spatiotemporal features S^e and further enhanced structural features P^e.

2.2.2. *Details of SpatioTemporal Evoformer*

In this subsection, we give further insights and details of STEvo. Its details are shown in the Fig. 1(b). We denote the branch processing S as the SpatioTemporal Modeling branch. And the branch for P is the Structural Enhancement branch.

SpatioTemporal Modeling. Following the original Evoformer, our STEvo also uses row- and column-wise self-attention (RSA and CSA) [10] to treat the MSA head input–pose features S. The RSA means the self-attention for each row of S. Note that S has the shape of $N \times J \times d_s$ and its row is a $J \times d_s$ matrix which means the human pose features in one frame. So the RSA is essential to the spatial message interaction and can be written as follows:

$$Attn_i = \text{Softmax}(Q_i K_i^T/\sqrt{d_s} + \text{Linear}(P))V_i, \quad (1)$$

where $Q_i \in \mathbb{R}^{J\times d_s}$, $K_i \in \mathbb{R}^{J\times d_s}$ and $V_i \in \mathbb{R}^{J\times d_s}$ denote query, key and value [8] respectively, which are computed by pose features $S_i \in \mathbb{R}^{J\times d_s}$ of the i-th frame. Compared to the vanilla transformer [8], RSA injects human priors P into the affinities to constrain relationships between different joints, which makes the spatial information interaction more reliable. The CSA is the vanilla self-attention for each column ($\mathbb{R}^{N\times d_s}$) of S, which is obviously corresponding to the temporal interaction for each joint. So the cascade of RSA and CSA can model the spatiotemporal patterns with human kinematic priors and lead to enhanced pose features S^e.

Structural Enhancement. The structural features P are updated by the enhanced pose features S^e. The outer product mean (OPM) [10] is first applied on different rows of S^e:

$$F_{ij} = \frac{1}{N}\sum_{n=0}^{N-1} S^e_{n,i} \otimes S^e_{n,j}, \tag{2}$$

where $S^e_{n,i} \in \mathbb{R}^{N\times d_s}$ denotes the features of the i-th joint at the n-th frame. The output $F_{ij} \in \mathbb{R}^{d_s\times d_s}$ means the temporal averaged outer product matrix along the whole N frames, indicating the high-order features of the relationship between the i-th and j-th joints. And then, these high-order features $F \in \mathbb{R}^{J\times J\times d_s\times d_s}$ are used to update P as follows:

$$\widetilde{P} = P + \text{Linear}(F). \tag{3}$$

This Linear$(\cdot)$ is a fully-connect layer operated on the channel level: $\mathbb{R}^{d_s\times d_s} \rightarrow \mathbb{R}^{d_p}$, which makes the dimension of F matched with P. After that, several non-attention and attention modules, including triangular updates using outgoing and incoming edges (OTU and ITU) and triangular self-attention around starting and ending node (STA and ETA) [10] which are utilized in the original Evoformer, are operated on $\widetilde{P}$ to get the final refined P^e. These refined structural features incorporate two kinds of human priors. One is extracted by the kinematic tree and the other is derived from the given pose sequence, leading to stronger human pose prior knowledge and improving the final 3D pose estimation accuracy.

2.3. Recursive Refinement

After obtaining the enhanced spatiotemporal and structural features S^e and P^e, the Recursive Refinement module estimates the 3D poses by the following recursive pipeline:

$$S^r_t = \text{FeatRe}(S^e_t, X_{t-1}), \;\; X_t = \text{RegHead}(S^r_t), \tag{4}$$

where X_t is the 3D pose estimation results at t-th round and $S^e_t = S^r_{t-1}$. Note that, $S^e_1 = S^e$, $X_0 = \mathbf{0}$, and the pipeline is shown in Fig. 1(c). The FeatRe refines the pose features S^e by introducing the last round pose estimation results X_{t-1}:

$$\begin{aligned} S^r_t &= \text{FeatRe}(S^e_t, X_{t-1}) \\ &= \underbrace{\text{RSA}(S^e_t, P^e) - \text{MLP}(X_{t-1})}_{\text{residual term}} + S^e_t. \end{aligned} \tag{5}$$

The residual term estimates the corresponding gap between S^e_t and X_{t-1}, reflecting new information that the X_{t-1} lack but S^e_t provide. The MLP$(\cdot)$ means a 3 layer multi-layer perception. This new information is added on the S^e_t to get the refined one S^r_t. The S^r_t has the shape of $N \times J \times d_s$, expressing each joint as a d_s-dimensional vector. So the RegHead is a joint-level convolutional module that maps each joint to its 3D coordinates. The last round estimation results are defined as $X^{3d} \in \mathbb{R}^{N\times J\times 3}$, storing the (x, y, z) coordinate value for each joint across all frames. The recursive estimation pipeline further injects the human priors in the final 3D pose estimation, leading to more reliable results.

2.4. Loss Function

The model is trained end-to-end with several types of loss functions. We first compute a coordinate loss $\mathcal{L}_c$ by the Mean Squared Error between estimated and ground truth 3D poses coordinates to constrain spatial error. Additionally, we compute the joint velocity loss [11] $\mathcal{L}_v$ and acceleration loss [12] $\mathcal{L}_a$, which measure the gap of the first-order and second-order derivative of the 3D pose coordinates, respectively. Moreover, we also project predicted 3D poses to 2D and compute the reprojection error between them and the ground truth 2D pose as the reprojection loss $\mathcal{L}_p$. Finally, these losses are combined together into a total one: $\mathcal{L} = \mathcal{L}_c + \lambda_v \cdot \mathcal{L}_v + \lambda_a \cdot \mathcal{L}_a + \lambda_p \cdot \mathcal{L}_p$, where $\lambda_\bullet$ is the weight of each loss term.

3. EXPERIMENTS

Implementation details. In our implementation, we set $\lambda_v = \lambda_a = 0.2$ and $\lambda_p = 0.1$. The RR module is conducted for $L = 2$ times. The proposed EvoPose is implemented by PyTorch on Tesla V100 GPU. We use Amsgrad optimizer with an initial learning rate of 0.001 which decays by 5% after each epoch and 50% after every 5 epochs. For fair comparisons, we also adopt horizontal flip augmentation following [6, 11, 15]. Note that, we only use the center frame of the final results as the prediction in inference.

3.1. Datasets and Evaluation Metrics

Human3.6M. The Human3.6M [16] is the largest indoor dataset for 3D HPE, containing 3.6 million video frames with 15 scenarios recorded by four calibrated cameras. Following [15, 6, 13, 14], we train our model on five subjects and test on two subjects with 17 joints. We report the mean per joint position error (MPJPE).

MPI-INF-3DHP. The MPI-INF-3DHP [17] is a large challenging 3D pose dataset with both indoor and outdoor scenarios. Following the settings in previous works [6, 13, 14], we use eight subjects for training and six subjects for testing on valid frames captured by eight cameras. We use three evaluation metrics: MPJPE, percentage of correct keypoints (PCK) under the threshold of 150mm, and area under curve (AUC).

3.2. Comparison with State-of-the-Art Methods

Results on Human3.6M. We compare our model with state-of-the-art methods. Cascaded pyramid network (CPN) [18] is used to estimate 2D poses from video frames, following [6, 13, 14, 11]. We report the results in Table 1 (top) with inputs of 243 frames and achieve **42.83mm** under MPJPE. It can be seen that our EvoPose is competitive to P-STMO [14]. To further explore the lower bound of our method, the results with ground truth 2D poses are reported in Table 1 (bottom). EvoPose surpasses all the other methods by a large margin, demonstrating the effectiveness of our method. When the input 2D poses are more accurate to be consistent with human kinematic, the performance of our method becomes better.

Results on MPI-INF-3DHP. To evaluate the ability of our model in the real world, we also report the performance on

Table 1: Results comparison with state-of-the-art methods on Human3.6M under MPJPE(mm). Top: 2D poses detected by CPN; Bottom: ground truth 2D poses. Bold: the best; Underline: the second.

Method	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
PoseFormer (ICCV'21)[6]	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
MHFormer (CVPR'22)[13]	39.2	43.1	40.1	40.9	44.9	51.2	**40.6**	41.3	53.5	60.3	**43.7**	**41.1**	43.8	29.8	30.6	43.0
P-STMO (ECCV'22)[14]	**38.9**	**42.7**	40.4	41.1	45.6	**49.7**	40.9	**39.9**	55.5	59.4	44.9	42.2	**42.7**	29.4	**29.4**	42.88
EvoPose (Ours)	41.7	43.0	**38.1**	**40.7**	**44.2**	52.5	41.3	42.6	**52.7**	**56.8**	45.3	41.5	42.9	**28.8**	29.6	**42.83**

Method	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
PoseFormer (ICCV'21)[6]	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
MHFormer (CVPR'22)[13]	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
P-STMO (ECCV'22)[14]	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
EvoPose (Ours)	**24.3**	**24.8**	**23.1**	**23.4**	**24.6**	**25.9**	**28.3**	**24.6**	**30.1**	**31.4**	**25.3**	**24.1**	**23.7**	**18.7**	**20.1**	**24.8**

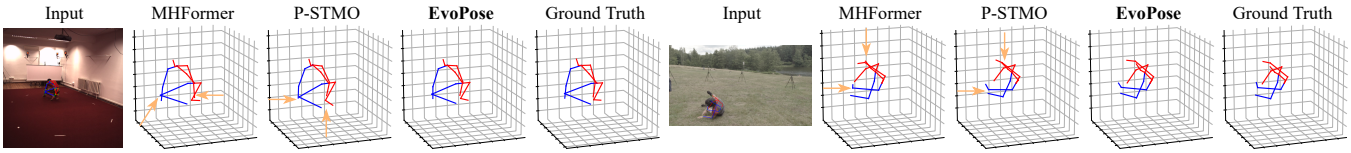

Fig. 2: Qualitative comparison with two state-of-the-art methods on both Human3.6M (left) and MPI-INF-3DHP (right) dataset.

Table 2: Results comparison with state-of-the-art methods on MPI-INF-3DHP. Bold: the best; Underline: the second.

Methods	N	PCK↑	AUC↑	MPJPE↓
PoseFormer (ICCV'21)[6]	9	88.6	56.4	77.1
MHFormer (CVPR'22)[13]	9	93.8	63.3	58.0
P-STMO (ECCV'22)[14]	81	**97.9**	75.8	32.2
EvoPose (Ours)	9	97.8	81.9	24.2
EvoPose (Ours)	27	97.8	**83.7**	**21.9**

Table 3: Ablation study on model components: SPR and the recursive pipeline in RR.

SPR	Recursive Pipeline (RR)	MPJPE
✗	✗	57.3
✓	✗	47.9
✗	✓	55.5
✓	✓	**45.8**

MPI-INF-3DHP compared to state-of-the-art methods. Following [15, 13, 14], we use ground truth 2D poses as inputs. Due to the sequence lengths of this dataset being shorter than Human3.6M, we adopt the setting of 27 frames. The results are shown in Table 2. It can be seen that our method performs much well compared to all the other methods with improvements of **10.3mm** in MPJPE and **7.9** in AUC over the previous best method, P-STMO, and achieves competitive performance in PCK with it by using smaller input length. This indicates the great power of our method in the real world.

Qualitative Results. We also show several visualization results in Fig. 2 on both Human3.6M and MPI-INF-3DHP with two state-of-the-art methods, MHFormer and P-SMTO. It is obvious that our model achieves excellent performance with more accurate angles on these two challenging datasets.

3.3. Ablation Study

The number of input frames is an important cause influencing the accuracy. We present the results with both CPN detected and ground truth 2D poses in Table 4. With the increase of the frame number, the results become more accurate.

Table 4: Ablation study on the number of input frames with MPJPE. CPN: cascaded pyramid network; GT: ground truth.

2D Inputs	9	27	81	243
CPN	48.3	45.8	43.7	**42.8**
GT	34.6	33.1	28.0	**24.8**

To further explore the effectiveness of components in our model, we conduct an additional ablation study on Human3.6M under MPJPE. Considering the time efficiency, we choose the number of 27 frames. As the results shown in Table 3, the structural features derived from the kinematic cues in SPR and enhanced in STE improve the accuracy by 9.4mm, while the reuse of estimated results can refine the performance to 55.5mm without the help of human priors. When the structural features get enhanced and work with RR to interact with estimated results iteratively, the performance is improved by a large margin of 11.5mm to 45.8mm. These indicate that the structural features effectively introduce human priors and can be better utilized by the recursive pipeline in RR to further improve the performance.

4. CONCLUSION

In this paper, we propose EvoPose, a novel recursive transformer for 3D HPE with kinematic structure priors. EvoPose first extracts structural features from the kinematic tree to introduce human priors and builds interactions between these features and 2D pose sequences to estimate 3D poses. Then EvoPose reuses previous estimations combined with the structural features to further utilize human priors for refinement. Extensive experiments show that the proposed EvoPose has a large performance advantage in real-world scenarios.

5. REFERENCES

[1] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos, "A review of 3d human pose estimation algorithms for markerless motion capture," *Computer Vision and Image Understanding*, vol. 212, pp. 103275, 2021.

[2] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang, "Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20502–20512.

[3] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *Acm transactions on graphics (tog)*, vol. 36, no. 4, pp. 1–14, 2017.

[4] Ching-Hang Chen and Deva Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 7035–7043.

[5] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang, "Exploiting temporal contexts with strided transformer for 3d human pose estimation," *IEEE Transactions on Multimedia*, 2022.

[6] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11656–11665.

[7] Mohammed Hassanin, Abdelwahed Khamiss, Mohammed Bennamoun, Farid Boussaid, and Ibrahim Radwan, "Crossformer: Cross spatio-temporal transformer for 3d human pose estimation," *arXiv preprint arXiv:2203.13387*, 2022.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Jogendra Nath Kundu, Siddharth Seth, MV Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty, "Kinematic-structure-preserved representation for unsupervised 3d human pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11312–11319.

[10] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al., "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[11] 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, pp. 7753–7762.

[12] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre, "3d human pose, shape and texture from low-resolution images and videos," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[13] 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, pp. 13147–13156.

[14] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao, "P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation," *arXiv preprint arXiv:2203.07628*, 2022.

[15] 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*, vol. 32, no. 1, pp. 198–209, 2021.

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[18] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 7103–7112.