Extended Abstract Track

ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

We provide theoretical and empirical evidence that, due to the depth ambiguity inherent to monocular 3D human pose estimation, traditional regression models suffer from pose-topology consistency issues, which standard evaluation metrics (MPJPE, P-MPJPE and PCK) fail to assess. We hence propose *ManiPose*, a manifold-constrained multi-hypothesis model for human-pose 2D-to-3D lifting. ManiPose addresses depth ambiguity by proposing multiple candidate 3D poses for each 2D input, each with its estimated plausibility. By constraining the outputs to lie on the human pose manifold, ManiPose guarantees the consistency of all hypothetical poses. We showcase the performance of ManiPose on simulated and real-world datasets, where it outperforms state-of-the-art models in pose consistency by a large margin while being very competitive on the MPJPE metric.

Keywords: human pose estimation, multiple choice learning, manifold estimation

1. Introduction

Monocular 3D human pose estimation (HPE) is a challenging learning problem that aims to predict 3D human poses given an image or a video from a single camera. Often, the problem is split into two successive steps: first 2D human pose estimation, then 2D-to-3D lifting. Due to depth ambiguity and occlusions, 2D-to-3D lifting is intrinsically ill-posed: multiple 3D poses correspond to the same projection observed in 2D. Despite that, the field has experienced fast developments, with substantial improvements in terms of mean-per-joint-prediction error (MPJPE) and derived metrics (*e.g.*, P-MPJPE, PCK) (Shan et al., 2023; Zhang et al., 2022; Zheng et al., 2021; Shan et al., 2022).

However, recent studies (Wehrbein et al., 2021; Holmquist and Wandt, 2023; Rommel et al., 2023) noted that poses predicted by state-of-the-art models fail to respect basic invariances of human morphology, such as bilateral sagittal symmetry, or constant distance between connected joints across time. In this work we address those issues and provide theoretical elements clarifying their cause. We show in particular that pose consistency and traditional performance metrics (such as MPJPE) cannot be optimized simultaneously by a standard regression model, because MPJPE ignores the topology of the space of human poses, and traditional regression models imply unimodality, thus overlooking the inherently ambiguous nature of 3D-HPE. We thus propose *ManiPose*, a novel approach for human-pose 2D-to-3D lifting which leverages multiple hypotheses and manifold constraints to address both depth ambiguity and pose consistency issues.

2. ManiPose

Following the previous state of the art, we split 3D-HPE into two steps, first estimating J human 2D keypoints in the pixel space from a sequence of T video frames $[\mathrm{x}_1, \ldots, \mathrm{x}_T] \in \mathbb{R}^{2\times J\times T}$, and then lifting them to 3D joint positions $[\hat{\mathrm{p}}_1, \ldots, \hat{\mathrm{p}}_T] \in \mathbb{R}^{3\times J\times T}$. We focus on the second step (*i.e.*, lifting) in the rest of the paper, assuming the availability of 2D keypoints x_i.

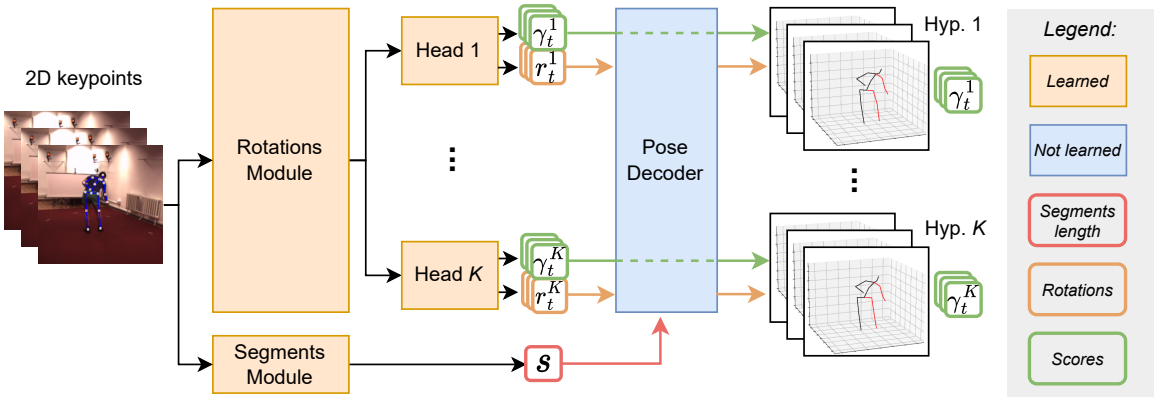


Figure 1: **Overview of ManiPose.** The rotations module predicts K possible sequences of segment rotations with their corresponding likelihoods (scores), while the segments module estimates the shared segment lengths.

2.1. Constraining predictions to the pose manifold

Human morphology prevents the joints from arbitrarily occupying the whole space. Instead, the poses within a movement are restricted to a manifold, reflecting the human skeleton's rigidity. If we knew the length of each segment connecting pairs of joints for a given subject, we could guarantee that the predicted poses lie on the correct pose manifold by only predicting the body part's rotations with respect to a reference skeleton. Since we do not have access to ground-truth segment lengths in real use cases, we propose to predict them, thus disentangling the estimation of the segments lengths (invariant across time) from the estimation of the joint rotations (variable across time).

We hence propose to use a neural network made of two parts (*cf.* Fig. 1):

1. the **segments module** predicts segment lengths $s \in \mathbb{R}^{J-1}$, shared by the T frames (time steps) of the input sequence;
2. the **rotations module** predicts the rotation $r = [r_{1,0}, \ldots, r_{T,J-1}] \in (\mathbb{R}^d)^{J \times T}$ of each joint relative to their parent joint at each time step. We represent rotations using 6D continuous embeddings (*i.e.*, $d = 6$) following insights from Zhou et al. (2019).

At each time t, disentangled representations s and r_t are decoded into poses $\hat{\mathrm{p}}_t \in (\mathbb{R}^3)^{J \times T}$ by applying the forward kinematics algorithm on a skeleton scaled according to s.

2.2. Multiple choice learning

As explained in the introduction, the inherent depth ambiguity of pose lifting requires multiple hypotheses to conciliate pose consistency and MPJPE performance. To address this, we adopt the multiple choice learning (MCL) framework (Lee et al., 2016), more precisely leveraging the *resilient MCL* approach proposed in Letzelter et al. (2024). Hence, instead of a single rotation $r_t \in (\mathbb{R}^d)^J$ per time step, ManiPose's rotations module is a multi-head network predicting K rotation hypotheses $r_t^k \in (\mathbb{R}^d)^J$ with corresponding likelihoods $\gamma_t^k \in [0, 1]$, called scores. Each one of the K rotation hypothesis is converted into a pose hypotheses using the same shared predicted segments length s (Fig. 1). As in Letzelter

et al. (2024), ManiPose is trained with a composite loss $\mathcal{L} = \mathcal{L}_{\text{wta}} + \beta\mathcal{L}_{\text{score}}$, made of a Winner-takes-all term $\mathcal{L}_{\text{wta}}$ and a scoring loss $\mathcal{L}_{\text{score}}$ (*cf.* appendix for details).

3. Formal analysis

ManiPose, as outlined in Section 2, is crafted to address the flaws inherent in unconstrained, single-hypothesis 3D-HPE. Hereafter, we formally highlight such limitations justifying our approach.

Let $\mathrm{p} = [\mathrm{p}^1, \dots, \mathrm{p}^J] \in \mathbb{R}^{3\times J}$ be a human pose, defined by the Cartesian 3D coordinates of each of the J joints of a predefined skeleton. Assuming bone length is fixed during a movement $\mathrm{m} = [\mathrm{p}_0, \dots, \mathrm{p}_T] \in \mathbb{R}^{3\times J\times T}$ of length T, then the poses p_t of m must all lie on the same manifold $\mathcal{M}$ of dimension $2(J-1)$ (homeomorphic to a product of spheres). This is proved and stated more precisely in the appendix. It implies that all poses predicted for a video sequence should ideally lie on the true manifold $\mathcal{M}$. We can show that minimizing joint position error using a single-hypothesis model necessarily violates this requirement:

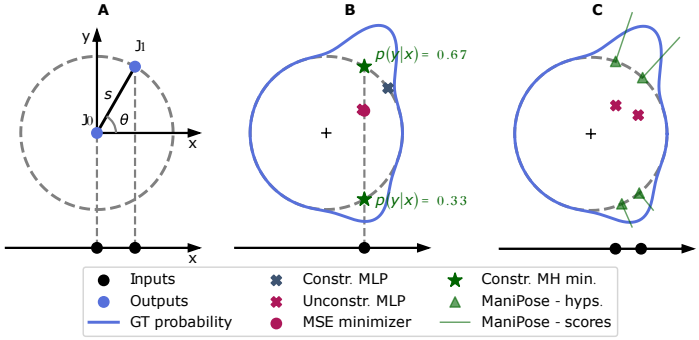


Figure 2: **(A)** 1D-to-2D articulated pose lifting problem. **(B)** True MSE minimizers under a multimodal distribution. One-to-one mappings cannot both reach optimal performance and stay on the pose manifold (dashed circle). **(C)** Multi-hypothesis approaches can deliver an acceptable solution to the problem.

Proposition 1 (Inconsistency of MSE minimizer) *With a rigid skeleton and mild assumptions on the training distribution, predicted 3D poses minimizing the traditional mean squared error (MSE) loss lie outside the pose manifold $\mathcal{M}$.*

Proof sketch. (See Appendix C.2). The conditional expectation $f^*(\mathrm{x}) = \mathbb{E}[\mathrm{p}\,|\,\mathrm{x}]$ is the minimizer of the MSE: $\mathbb{E}_{\mathrm{x},\mathrm{p}}\left[\|\mathrm{p} - f(\mathrm{x})\|_2^2\right]$. Hence, as the functions $(\ell_j)_{j=1}^{J-1}$ computing the lengths of the segments in a pose are stricly convex, Jensen's inequality leads to $\ell_j^2\left(f^*(\mathrm{x})\right) < s_j^2$, where s_j is the true length of the segment associated with joint j. ■

Proposition 1 has the following implications:

1. Traditional unconstrained single-hypothesis approaches are bound to predict inconsistent movements, where bone lengths may vary.
2. With a single hypothesis, models constrained to the manifold will always lose to unconstrained models in terms of MPJPE performance (*cf.* Corollary 8).

Table 1: **1D-to-2D performance.**

	MPJPE ↓	Distance to circle ↓
×	0.748	0.411
×	0.759	**0.000**
▲	**0.733**	**0.000**

We show next that multiple hypotheses allows to conciliate these antagonistic objectives.

4. Experiments

4.1. Insights to the formal argument on a simplified setting

We illustrate the argument of Section 3 in a simplified 1D-to-2D lifting setup with 2 joints, as depicted on Fig. 2-A. We train three different models with comparable architectures on

two datasets $\{(x_i, (x_i, y_i))\}_{i=1}^{N}$ sampled from the angular distributions represented in blue on Fig. 2-B,C. The models correspond to: (×) a 2-layer MLP trained to minimize the MSE between true and predicted joint positions (x, y) ; (×) an MLP of the same size constrained to the manifold, *i.e.* , the circle ; and (▲) our constrained multi-hypothesis model using the same MLP backbone, capable of predicting $K = 2$ poses.

Fig. 2 shows that the traditional unconstrained single-hypothesis approach (×) fails when facing a bimodal distribution (C), leading to predictions outside the circle, as depth ambiguity makes the lifting problem ill-posed. The single-hypothesis constrained model (×) delivers predictions on the circle, at the cost of worse MPJPE performance than the unconstrained MLP (Table 1). Such performance decrease is due to the Euclidean topology of the MPJPE metric having its minimum (●) outside the manifold (Fig. 2-B). Predicting multiple hypotheses constrained to the circle (★ in Fig. 2-B) allows escaping this dilemma, which is exactly what ManiPose does (▲ in Fig. 2-D). Those advantages translate into perfect pose consistency, and improved MPJPE performance (Table 1).

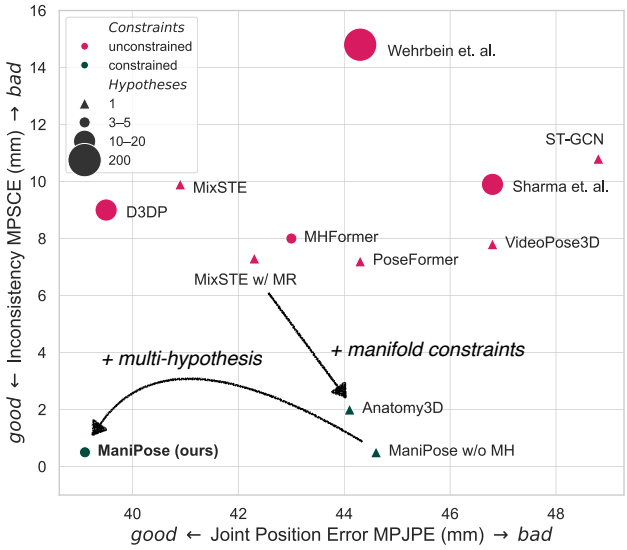


Figure 3: **Optimizing both 3D position and pose consistency requires combining constraints and multiple hypotheses.**

4.2. Comparison with state-of-the-art

To confirm our insights from the toy problem, we compare ManiPose to state-of-the-art 3D-HPE models on the well-known Human 3.6M dataset (Ionescu et al., 2014), using the MixSTE model as our backbone and 5 hypothesis heads (see further details in appendix). Fig. 3 shows that ManiPose outperforms previous methods in terms of MPJPE, while reaching nearly perfect consistency. Moreover, note that MPJPE and consistency metrics are not positively correlated for single-hypothesis methods. As predicted in Section 3, we see that MPJPE improvements achieved by previous unconstrained single and multi-hypothesis methods usually come at the cost of poorer consistency. Training and architecture details, metrics expressions and additional result tables and ablations can be found in the appendix.

5. Conclusion

We presented a new manifold-constrained multi-hypothesis human pose lifting method (ManiPose) and demonstrated its empirical superiority to the existing state-of-the-art in terms of manifold consistency and traditional metrics. Further, we provided theoretical evidence supporting the tenets of our method, by proving the inherent limitation of existing 3D-HPE approaches and exposing them in a simplified setting.

References

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*, pages 2272–2281, 2019.

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*, 32(1):198–209, 2021.

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*, pages 7103–7112, 2018.

Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15977–15987, 2023.

Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 68–84, 2018.

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*, 36(7): 1325–1339, July 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. In *International Conference on Machine Learning*, pages 2014–2023. PMLR, 2017.

Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems*, 29, 2016.

Victor Letzelter, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gael Richard. Resilient multiple choice learning: A learned scoring scheme with application to audio scene analysis. *Advances in neural information processing systems*, 36, 2024.

Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019.

Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. In *British Machine Vision Conference (BMVC)*, 2020.

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*, pages 3383–3393, 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*, pages 13147–13156, 2022.

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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017.

Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.

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*, pages 7753–7762, 2019.

Cédric Rommel, Eduardo Valle, Mickaël Chen, Souhaiel Khalfaoui, Renaud Marlet, Matthieu Cord, and Patrick Pérez. DiffHPE: Robust, Coherent 3D Human Pose Lifting with Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3220–3229, 2023.

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, July 2022.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023.

Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2325–2334, 2019.

Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6349–6357, 2019.

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021.

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*, pages 13232–13242, 2022.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D Human Pose Estimation with Spatial and Temporal Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11636–11645, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01145.

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.

Appendix A. Further details about the method

A.1. Constraining predictions to the pose manifold

Rationale. Human morphology prevents the joints from arbitrarily occupying the whole space. Instead, the poses within a movement are restricted to a manifold, reflecting the human skeleton's rigidity. If we knew the length of each segment connecting pairs of joints for a given subject, we could guarantee that the predicted poses lie on the correct pose manifold by only predicting the body part's rotations with respect to a reference skeleton. Since we do not have access to ground-truth segment lengths in real use cases, we propose to predict them, thus disentangling the estimation of the reference lengths (fixed across time) from the estimation of the joint rotations (variable across time).

Disentangled representations. We constrain model predictions to lie on an estimated manifold by predicting parametrized disentangled transformations of a reference pose $\mathrm{u} \in (\mathbb{R}^3)^J$, for which all segments have unit length. Namely, we propose to split the network into two parts (*cf.* Fig. 1):

1. **Segments module**, which predicts segment lengths $s \in \mathbb{R}^{J-1}$, shared by the T frames (time steps) of the input sequence;
2. **Rotations module**, which predicts the rotation $r = [r_{1,0}, \ldots, r_{T,J-1}] \in (\mathbb{R}^d)^{J \times T}$ of each joint relative to their parent joint at each time step.

Rotations representation. We represent rotations using 6D continuous embeddings (*i.e.*, $d = 6$). Compared to quaternions or axis-angles, those representations are continuous and, hence, better learned by neural networks, as demonstrated by their proposers (Zhou et al., 2019).

Pose decoding. To deliver pose predictions in $(\mathbb{R}^3)^{J \times T}$, the intermediate representations (s, r) must be decoded. We achieve that in three steps (*cf.* Fig. 4):

1. We scale the unit segments of the reference pose $\mathrm{u} \in (\mathbb{R}^3)^J$ using s, forming a scaled reference pose u': $\mathrm{u}'_j = \mathrm{u}'_{\tau(j)} + s_j(\mathrm{u}_j - \mathrm{u}_{\tau(j)})$ for $0 < j \leq J-1$, where τ maps the index of a joint to its parent's, if any.
2. For each time step $1 \leq t \leq T$ and joint $0 \leq j < J$, we convert the predicted rotation representations $r_{t,j}$ into rotation matrices $R_{t,j} \in \mathrm{SO}(3)$ (Algorithm 1).
3. We apply those rotation matrices $R_{t,j}$ at each time step t to the scaled reference pose u' using forward kinematics (Algorithm 2).

A.2. Multiple choice learning

ManiPose architecture. As explained in the introduction, the inherent depth ambiguity of pose lifting requires multiple hypotheses to conciliate pose consistency and MPJPE performance. To address this, we adopt the multiple choice learning (MCL) (Lee et al., 2016) framework, more precisely leveraging the *resilient MCL* approach as proposed by Letzelter et al. (2024). This methodology allows the estimation of conditional distributions for regression tasks, enabling our model to predict multiple plausible 3D poses for each

2D input. Specifically, instead of a single rotation $r_t \in (\mathbb{R}^d)^J$ per time step, ManiPose's rotations module predicts an intermediate representation $e_t \in (\mathbb{R}^{d'})^J$ that feeds K linear heads (with weights W_r^k and W_γ^k), each predicting its own rotation hypothesis $r_t^k \in (\mathbb{R}^d)^J$ with a corresponding likelihood $\gamma_t^k \in [0,1]$. That is, for all $1 \le t \le T$, $r_t^k = W_r^k e_t$ and $\gamma_t^k = \sigma[\tilde{\gamma}_t]_k$, where the softmax function σ is applied to the vector $\tilde{\gamma}_t = [\tilde{\gamma}_t^1, \dots, \tilde{\gamma}_t^K] \in \mathbb{R}^K$ of intermediate values $\tilde{\gamma}_t^k = W_\gamma^k e_t$.

All rotation hypotheses are decoded together with the shared segment-length predictions s, resulting in K hypothetical pose sequences $\hat{\mathrm{p}}^k = (\hat{\mathrm{p}}_t^k)_{t=1}^T$, with corresponding likelihood sequences $\gamma^k = (\gamma_t^k)_{t=1}^T$, called **scores** hereafter (Fig. 1).

Loss function. As in Letzelter et al. (2024), ManiPose is trained with a composite loss

$$\mathcal{L} = \mathcal{L}_{\text{wta}} + \beta \mathcal{L}_{\text{score}} \,. \tag{1}$$

The first term, $\mathcal{L}_{\text{wta}}$, is the winner-takes-all loss (Lee et al., 2016)

$$\mathcal{L}_{\text{wta}}(\hat{\mathrm{p}}(\mathrm{x}), \mathrm{p}) = \frac{1}{T} \sum_{t=1}^{T} \min_{k \in [\![1,K]\!]} \ell(\hat{\mathrm{p}}_t^k(\mathrm{x}), \mathrm{p}_t) \,, \tag{2}$$

where $\ell(\hat{\mathrm{p}}_t^k(\mathrm{x}), \mathrm{p}_t) \triangleq \frac{1}{J} \sum_{j=0}^{J-1} \|\mathrm{p}_{t,j} - \hat{\mathrm{p}}_{t,j}^k(\mathrm{x})\|_2$, and $\hat{\mathrm{p}}_t^k(\mathrm{x})$ denotes the pose prediction at time t using the k^{th} head. The second term, $\mathcal{L}_{\text{score}}$, is the scoring loss

$$\mathcal{L}_{\text{score}}(\hat{\mathrm{p}}(\mathrm{x}), \gamma(\mathrm{x}), \mathrm{p}) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{H}\big(\delta(\hat{\mathrm{p}}_t, \mathrm{p}_t), \gamma_t(\mathrm{x})\big) \,, \tag{3}$$

where $\mathcal{H}(\cdot,\cdot)$ is the cross-entropy, $\hat{\mathrm{p}}_t = (\hat{\mathrm{p}}_t^k)_{k=1}^K$, and

$$[\delta(\hat{\mathrm{p}}_t, \mathrm{p}_t)]_k \triangleq \mathbf{1}\Big[k \in \operatorname*{arg\,min}_{k' \in [\![1,K]\!]} \ell\left(\hat{\mathrm{p}}_t^{k'}, \mathrm{p}_t\right)\Big] \tag{4}$$

is the indicator function of the *winner* pose hypothesis, which is the closest to the ground truth. Eq. (3) is the average cross-entropy between target and predicted scores $\gamma_t(\mathrm{x}) \in [0,1]^K$ at each time t.

Those losses are complementary. The winner-takes-all loss updates only the best predicted hypothesis, specializing each head on part of the data distribution (Lee et al., 2016). The scoring loss allows the model to learn how likely each head is to winning, thus avoiding overconfidence of non-winner heads (*cf.* Lee et al. (2017); Tian et al. (2019)).

Conditional distribution estimation. As detailed in Letzelter et al. (2024), the model may be interpreted probabilistically as a multimodal conditional density estimator. More precisely, it models the distribution $\mathrm{P}(\mathrm{p}|\mathrm{x})$ of 3D poses conditioned on 2D poses as a mixture of Dirac distributions:

$$\hat{\mathrm{P}}(\mathrm{p}|\mathrm{x}) \triangleq \sum_{k=1}^{K} \gamma^k(\mathrm{x}) \delta_{\hat{\mathrm{p}}^k(\mathrm{x})}(\mathrm{p}) \,. \tag{5}$$

Hence, the predicted conditional distribution has, at each predicted hypothesis $\hat{\mathrm{p}}^k$, a peak whose likelihood is given by the predicted score γ^k. As described in Section 3, interpreting hypotheses and scores probabilistically enables us to handle depth ambiguity.

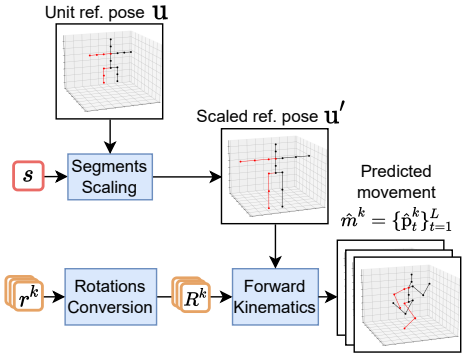


Figure 4: **Pose decoder overview.**

Appendix B. Further experimental details and results

B.1. Experimental setup

Datasets. We evaluate our model on two 3D-HPE datasets. **Human 3.6M** (Ionescu et al., 2014) contains 3.6 million images of 7 actors performing 15 different indoor actions. It is the most widely used dataset for 3D-HPE. Following previous works (Zhang et al., 2022; Li et al., 2022; Zheng et al., 2021; Pavllo et al., 2019), we train on subjects S1, S5, S6, S7, S8, and test on subjects S9 and S11, adopting a 17-joint skeleton (*cf.* Fig. 5). We employ a pre-trained CPN (Chen et al., 2018) to compute the input 2D keypoints, as in Pavllo et al. (2019); Zhang et al. (2022). **MPI-INF-3DHP** (Mehta et al., 2017) also adopts a 17-joint skeleton, but, with fewer samples and containing both indoor and outdoor scenes, it is more challenging than Human 3.6M. We used ground-truth 2D keypoints for this dataset, as usually done (Zheng et al., 2021; Chen et al., 2021; Zhang et al., 2022).

Traditional evaluation metrics. The mean per-joint position error (MPJPE) is the usual performance metric for the datasets above, under different protocols, both reported in mm. In protocol #1, the root joint position is set as a reference, and the predicted root position is translated to 0. In protocol #2 (P-MPJPE), predictions are additionally Procrustes-corrected. For MPI-INF-3DHP, additional thresholded metrics derived from MPJPE are often reported, such as AUC (Area Under Curve) and PCK (Percentage of Correct Keypoints) with a threshold at 150 mm, as explained in Mehta et al. (2017).

Pose consistency metrics. MPJPE being insufficient to assess pose consistency (Section 3), we further assess to which extent predicted skeletons are rigid by measuring the average standard deviations of segment lengths across time in predicted action sequences:

$$\text{MPSCE} \triangleq \frac{1}{J-1}\sum_{j=1}^{J-1}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(s_{t,j,\tau(j)} - \bar{s}_{j,\tau(j)})^2}\,, \tag{6}$$

with $s_{t,j,i} = \|\hat{\mathrm{p}}_{t,j} - \hat{\mathrm{p}}_{t,i}\|_2$ and $\bar{s}_{j,i} = \frac{1}{T}\sum_{t=1}^{T} s_{t,j,i}$, where τ was defined in Section 2.1. We call this metric, reported in mm, the Mean Per Segment Consistency Error (MPSCE).

Following Holmquist and Wandt (2023); Rommel et al. (2023), we also assess the bilateral symmetry of predicted skeletons through the Mean Per Segment Symmetry Error

(MPSSE), in mm:

$$\text{MPSSE} \triangleq \frac{1}{T\,|\mathcal{J}_{\text{left}}|} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}_{\text{left}}} \left| s_{t,j,\tau(j)} - s_{t,j',\tau(j')} \right|, \qquad \text{with} \quad j' = \zeta(j)\,, \tag{7}$$

where $\mathcal{J}_{\text{left}}$ denotes the set of indices of left-side joints and ζ maps left-side joint indices to their right-side counterparts.

Multi-hypothesis evaluation protocol. One must decide how to use multiple hypotheses to compute the metrics. The dominant approach (Li and Lee, 2019, 2020; Oikarinen et al., 2021; Sharma et al., 2019; Wehrbein et al., 2021; Holmquist and Wandt, 2023) is the **oracle** evaluation, *i.e.*, using the predicted hypothesis closer to the ground truth (*i.e.*, Eq. (2) for MPJPE). That makes sense for multi-hypothesis methods, as the oracle metric measures the distance between the target and the discrete set of predicted hypotheses. It aligns with the idea of many possible outputs for a given input.

Hypotheses can also be *aggregated* into a final pose, *e.g.*, through unweighted or weighted averaging (using predicted scores). The latter has the disadvantage of falling back to a one-to-one mapping scheme, which is precisely what we want to avoid in a multi-hypothesis setting.

We report both oracle and aggregated metrics in our experiments, favoring oracle results.

Implementation details. ManiPose, as presented in Section 2, is compatible with any backbone. Here, we chose to build on the MixSTE (Zhang et al., 2022) network for both the rotations and the segment modules (the latter in a reduced scale). Details about our architecture and training appear in Appendix E.

B.2. Comparison with the state of the art

Table 2: **Pose consistency evaluation of state-of-the-art methods on Human3.6M.** MPJPE performance and pose consistency are not correlated; only ManiPose excels in both. T: sequence length. K: number of hypotheses. Orac.: Metric computed using oracle hypothesis. **Bold**: best; Underlined: second best. *: MPSSE values reported in Holmquist and Wandt (2023). Missing entries: methods with unavailable code.

	T	K	Orac.	MPJPE↓	MPSSE↓	MPSCE↓
Single-hypothesis methods:						
ST-GCN (Cai et al., 2019)	7	1		48.8	8.9	10.8
VideoPose3D (Pavllo et al., 2019)	243	1		46.8	6.5	7.8
PoseFormer (Zheng et al., 2021)	81	1		44.3	4.3	7.2
Anatomy3D (Chen et al., 2021)	243	1		44.1	1.4	2.0
MixSTE (Zhang et al., 2022)	243	1		40.9	8.8	9.9
Multi-hypothesis methods:						
Sharma (Sharma et al., 2019)	1	10	✓	46.8	13.0	9.9
Wehrbein (Wehrbein et al., 2021)	1	200	✓	44.3	12.2	14.8
Diffpose (Holmquist and Wandt, 2023)*	1	200	✓	43.3	14.9	-
MHFormer (Li et al., 2022)	351	3		43.0	5.7	8.0
D3DP (P-Best) (Shan et al., 2023)	243	20	✓	39.5	6.9	9.0
ManiPose (Ours)	243	5	✓	**39.1**	**0.3**	**0.5**

Human 3.6M. Comparisons with state-of-the-art single- and multi-hypothesis methods are presented in Table 2 and illustrated in Fig. 3. ManiPose outperforms previous methods in

terms of MPJPE, while reaching nearly perfect consistency. Moreover, note that MPJPE and consistency metrics are not positively correlated for single-hypothesis methods. As predicted in **??**, our empirical results show that MPJPE improvements achieved by MixSTE come at the cost of poorer consistency compared to previous models. In contrast, the only single-hypothesis constrained model, Anatomy3D (Chen et al., 2021), achieves good consistency at the expense of inferior MPJPE. Those results empirically validate the theoretical predictions of **??**, further confirming what we have shown, intuitively, in the simplified 1D-to-2D setting (Section 4.1). Note that while ManiPose is deterministic, previous multi-hypothesis methods are generative and frame-based, except for MHFormer. Table 2 shows that they require up to two orders of magnitude more hypotheses than ManiPose to reach competitive performance. More detailed MPJPE results per action appear in **????** in the supplemental.

Fig. 6 showcases qualitative results, where multiple hypotheses help in depth-ambiguous situations.

Table 3: **Comparison with the state-of-the-art on MPI-INF-3DHP using ground-truth 2D poses.** T: sequence length.

	T	PCK↑	AUC↑	MPJPE↓	MPSSE↓	MPSCE↓
VideoPose3D (Pavllo et al., 2019)	81	85.5	51.5	84.8	10.4	27.5
PoseFormer (Zheng et al., 2021)	9	86.6	56.4	77.1	10.8	14.2
MixSTE (Zhang et al., 2022)	27	94.4	66.5	54.9	17.3	21.6
P-STMO (Shan et al., 2022)	81	97.9	<u>75.8</u>	**32.2**	<u>8.5</u>	<u>11.3</u>
ManiPose (Ours) Aggr.	27	<u>98.0</u>	75.3	37.7	**0.6**	**1.3**
ManiPose (Ours) Orac.	27	**98.4**	**77.0**	<u>34.6</u>	**0.6**	**1.3**

MPI-INF-3DHP. Similar results were obtained for this dataset (*cf.* Table 3). Not only does ManiPose reach consistency errors close to 0, but also best PCK and AUC performance. As for MPJPE, only Shan et al. (2022) achieves slightly better performance, at the cost of large pose consistency errors.

B.3. Ablation study

Table 4: **Ablation study: Single hypothesis cannot optimize both MPJPE and consistency.** ManiPose uses the same backbone as MixSTE. MR: with manifold regularization. MC: manifold-constrained. **Bold**: best. <u>Underlined</u>: second best.

	MR	MC	K	# Params.	MPJPE↓	MPSSE↓	MPSCE↓
ManiPose (Ours)	✗	✓	5	34.44 M	**39.1**	**0.3**	**0.5**
w/o MH	✗	✓	1	34.42 M	44.6	**0.3**	**0.5**
w/o MC, w/ MR	✓	✗	1	33.78 M	42.3	<u>5.7</u>	<u>7.3</u>
w/o MR (MixSTE)	✗	✗	1	33.78 M	<u>40.9</u>	8.8	9.9

Impact of components. We evaluate the impact of removing each component of ManiPose on the Human 3.6M performance (Table 4). The components tested are the multiple hypotheses (MH) and the manifold constraint (MC). We also compare MC to a more standard manifold regularization (MR), *i.e.*, adding Eq. (6) to the loss. Note that without all these components, we fall back to MixSTE (Zhang et al., 2022), and that the performances reported in Table 4 also appear in Fig. 3.

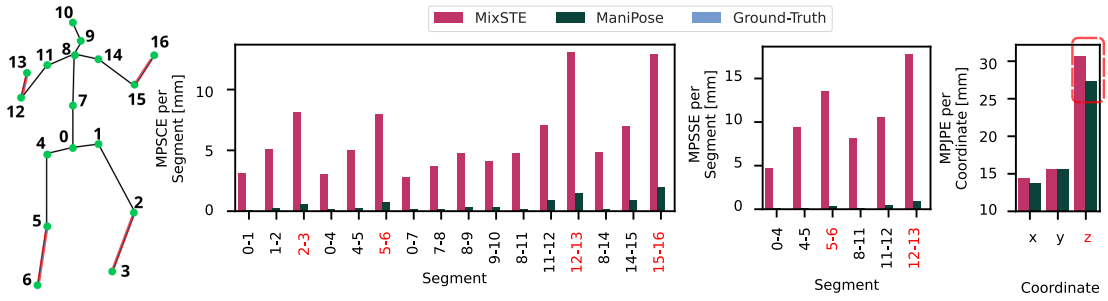


Figure 5: **MPSCE, MPSSE and MPJPE per segment/coordinate (lower is better).** ManiPose mostly helps to deal with the depth ambiguity (z coordinate). Ground-truth poses are represented but not visible because they have perfect consistency.

We see that MR helps to improve pose consistency, but not as much as MC. However, without multiple hypotheses, MC consistency improvements come at the cost of degraded MPJPE performance, as foreseen by our formal analysis (Section 3). Only the combination of both MC and MH allows us to optimize both consistency and MPJPE.

Fine error analysis. We can see in Fig. 5 that, compared to MixSTE, ManiPose reaches substantially superior MPSSE and MPSCE, consistency across all skeleton segments. Furthermore, note that larger MixSTE errors occur for segments KNEE-FOOT and ELBOW-WRIST, which are the most prone to depth ambiguity. That agrees with coordinate-wise errors depicted in Fig. 5, showing that ManiPose improvements mostly translate into a reduction of MixSTE depth errors, which are twice as large as for other coordinates. Further ablations, including the effect of the number of hypotheses K and the score loss weight β appear in the supplemental.

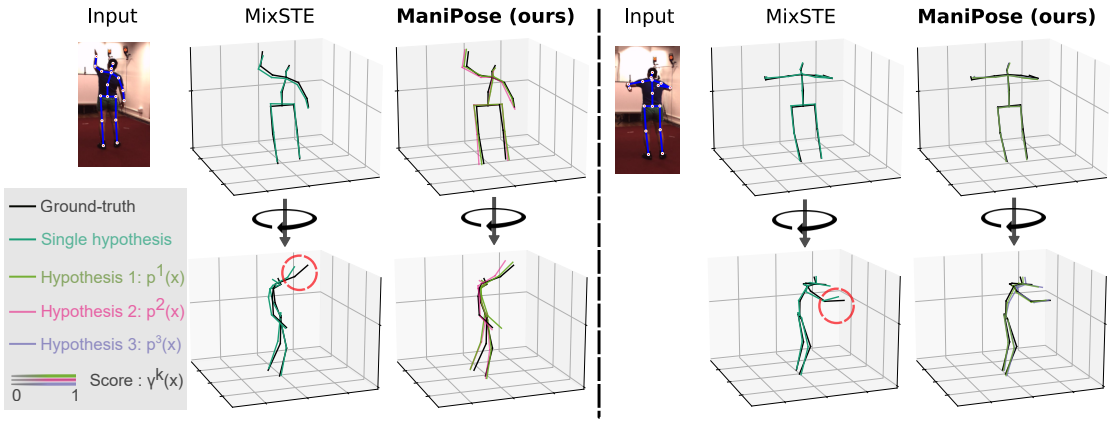


Figure 6: **Qualitative comparison between ManiPose and state-of-the-art regression method, MixSTE.** Two pairs of predicted hypotheses by ManiPose are illustrated in green-pink (left) and green-purple (right), where opacity is used to represent the predicted scores. Multiple hypotheses and constraints help to deal with depth ambiguities and avoids predicting shorter limbs (red circles).

Appendix C. Further theoretical elements

C.1. Assumptions verifications

Let us first define a few elements that we will need needed for our derivations.

Definition 2 (Human skeleton) *We define a human skeleton as an undirected connected graph $G = (V, E)$ with $J = |V|$ nodes, called* joints*, associated with different human body articulation points. We assume a predefined order of joints and denote $A = [A_{ij}]_{0 \leq i,j < J} \in \{0,1\}^{J \times J}$ the adjacency matrix of G, defining joints connections.*

Definition 3 (Human pose and movement) *Let G be a skeleton of J joints. We attach to each joint i a position p_i^G in $\mathbb{R}^3$ and call the vector $\mathrm{p}^G = [\mathrm{p}_0^G, \dots, \mathrm{p}_{J-1}^G] \in (\mathbb{R}^3)^J$ a* human pose*. Furthermore, given a series of increasing time steps $t_1 < t_2 < \cdots < t_T \in \mathbb{R}$, we define a human* movement m *as a sequence of poses* of the same subject *at those instants $\mathrm{m} = [\mathrm{p}_{t_1}^G, \dots, \mathrm{p}_{t_T}^G] \in (\mathbb{R}^3)^{J \times T}$.*

We base the theoretical results of **??** on the following assumptions. The first states the reference frame traditionally used for assessing 3D-HPE models:

Assumption 4 (Reference root joint) *For any skeleton G and movement m of length T, the joint of index 0, called the* root joint*, is at the origin $\mathrm{p}_{t,0}^G = [0,0,0]$ at all times $t_1 \leq t \leq t_T$. That is equivalent to measuring positions p_t^G in a reference frame attached to the root joint.*

The second assumption concerns the rigidity of human body parts:

Assumption 5 (Rigid segments) *We assume that the Euclidean distance between adjacent joints is constant within a movement m: for any pair of instants t and t' and for any joints i, j such that $A_{ij} = 1$, we assume that*

$$s_{t,i,j} = s_{t',i,j} = s_{i,j}\,, \tag{8}$$

where $s_{t,i,j} = \|\mathrm{p}_{t,i}^G - \mathrm{p}_{t,j}^G\|_2 > 0$.

Finally, we assume that the conditional distribution of poses does not collapse to a single point, *i.e.*, that we have a one-to-many problem:

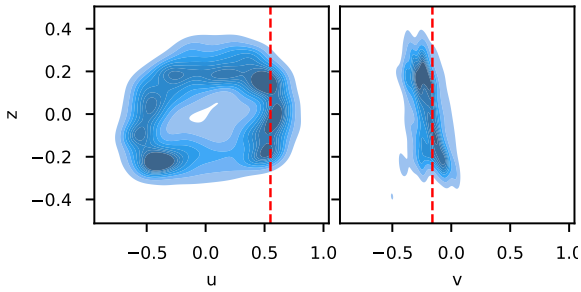
Assumption 6 (Non-degenerate conditional distribution) *Given a joint distribution $\mathrm{P}(\mathrm{x}^G, \mathrm{p}^G)$ of 3D poses $\mathrm{p}^G \in (\mathbb{R}^3)^J$ and corresponding 2D inputs $\mathrm{x}^G \in (\mathbb{R}^2)^J$, we assume that the conditional distribution $\mathrm{P}(\mathrm{p}^G|\mathrm{x}^G)$ is non-degenerate,* i.e.*, it is not a single Dirac distribution.*

Note that can be true even when $\mathrm{P}(\mathrm{x}^G, \mathrm{p}^G)$ is unimodal (*e.g.*, Fig. 2).

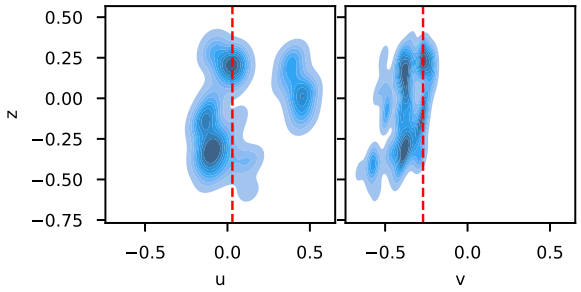
We verified on Human 3.6M (Ionescu et al., 2014) ground-truth data that assumptions 5 and 6 hold for actual poses in both training and test splits.

Segments rigidity. As shown on Fig. 5, ground-truth 3D poses have perfect MPSSE (7) and MPSCE (6) metrics, meaning that ground-truth skeletons are perfectly symmetric, with rigid segments. Assumption 5 is thus verified in actual training and test data.

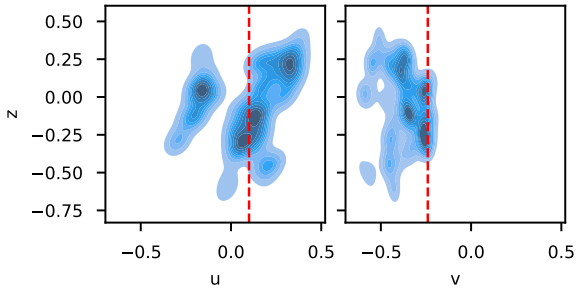
Non-degenerate distributions. As shown on Fig. 7, the conditional distribution of ground-truth 3D poses given 2D keypoints position is clearly multimodal, and, thus, non-degenerate (not reduced to a single Dirac distribution). That validates assumption 6 and explains why multi-hypothesis techniques are necessary.



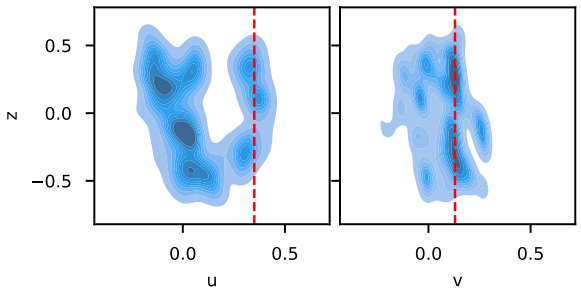
(*a*) S9, Walking



(*b*) S1, Greeting



(*c*) S11, Directions



(*d*) S1, SittingDown

Figure 7: **Estimated joint distributions of ground-truth 2D inputs (u, v pixel coordinates) together with 3D z-coordinates (depth) for different subjects and actions.** The depth density conditional on inputs is clearly multimodal. Vertical red lines are examples of depth-ambiguous inputs. Distributions are estimated with a kernel density estimator from the Seaborn plotting library (Waskom, 2021).

C.2. Proofs and additional corollaries

This section contains the proofs of the theoretical results presented in Section 3, together with a few corollaries.

First, we state more formally our result on the existence of a human pose manifold:

Proposition 7 (Human pose manifold) *Assuming a rigid skeleton (assumptions 4 and 5), all poses of a movement* $\mathrm{m} = [\mathrm{p}_t]_{t=1}^T$ *lie on a manifold* $\mathcal{M}$ *of dimension* $2(J-1)$*:*

$$\forall t \in \{1, \dots, T\}, \quad \mathrm{p}_t \in \mathcal{M}\,. \tag{9}$$

Proof [Proposition 7] Let i be a joint connected to the root p_0 (*i.e.*, $A_{i0} = 1$). From assumptions 4 and 5, we know that at any instant t, $\mathrm{p}_{t,i}^G$ lies on the sphere $S^2(0, s_{i,0})$ centered at 0 with radius $s_{i,0}$ independent of time. Therefore, its position can be fully parameterized in spherical coordinates by two angles $(\theta_{t,i}, \phi_{t,i})$. Let j be a joint connected to i. Like before, assumption 5 implies that at any instant t, $\mathrm{p}_{t,j}^G$ lies on the moving sphere $S^2(p_{t,i}^G, s_{j,i})$ centered at $p_{t,i}^G$ with radius $s_{j,i}$ independent of time. Thus, we can fully describe $\mathrm{p}_{t,j}^G$ with the position of its center, $\mathrm{p}_{t,i}^G$ and the spherical coordinates $(\theta_{t,j}, \phi_{t,j})$ of joint j relative to the center of the sphere, *i.e.*, joint i. That means that there is a bijection between the possible positions attainable by $\mathrm{p}_{t,j}^G$ at any instant and the direct product of spheres $S^2(0, s_{i,0}) \otimes S^2(0, s_{j,i})$.[1] That bijection is an homeomorphism since it is a composition of homeomorphisms: we can compute $\mathrm{p}_{t,j}^G$ from $(\theta_{t,i}, \phi_{t,i}, \theta_{t,j}, \phi_{t,j})$ following the forward kinematics algorithm (Murray et al., 2017) (*cf.* algo. 2), *i.e.*, using a composition of rotations and translations.

Now let us assume for some arbitrary joint k that $\mathrm{p}_{t,k}^G$ lies at all times on a space $\mathcal{M}_{2d}$ homeomorphic to a product of spheres of dimension $2d$. That means that $\mathrm{p}_{t,k}^G$ can be fully parametrized using $2d$ spherical angles $(\theta_1, \phi_1, \dots, \theta_d, \phi_d)$. Let l be a joint connected to k (typically one further step away from the root joint p_0 and not already represented in $\mathcal{M}_{2d}$). As before, at any instant t, $\mathrm{p}_{t,l}^G$ needs to lie on the sphere centered on $\mathrm{p}_{t,k}^G$ of constant radius $s_{k,l}$. Thus, we can fully describe $\mathrm{p}_{t,l}^G$ using the $2(d+1)$-tuple of angles obtained by concatenating its spherical coordinates relative to joint k, together with the $2d$-tuple describing $\mathrm{p}_{t,k}^G$, *i.e.* the center of the sphere. So $\mathrm{p}_{t,l}^G$ lies on a space $\mathcal{M}_{2(d+1)}$ homeomorphic to a product of spheres of dimension $2(d+1)$.

We can conclude by induction that at any instant t, $\mathrm{p}_t = [\mathrm{p}_{t,1}^G, \dots, \mathrm{p}_{t,J}^G]$ lies on the same subspace of $(\mathbb{R}^3)^J$, which is homeomorphic to a product of spheres centered at the origin:

$$\bigotimes_{i<j/A_{ij}=1} S^2(0, s_{i,j})\,. \tag{10}$$

Finally, the previous space is trivially homeomorphic to $(S^2)^{J-1}$ through the scaling $(1/s_{i,j})_{i<j/A_{ij}=1}$. $(S^2)^{J-1}$ is a manifold of dimension $2(J-1)$ as the direct product of $J-1$ manifolds of dimension 2. ■ ■

Proof [Proposition 1] Let G be a skeleton with J joints, $\mathrm{x} \in (\mathbb{R}^2)^J$ a 2D pose, $\mathrm{p} \in (\mathbb{R}^3)^J$ its corresponding 3D pose, and $\mathrm{P}(\mathrm{x}, \mathrm{p})$ a joint distribution of poses in 2D and 3D. We define

1. $S^2(0, s_{j,i})$ is homeomorphic to $S^2(\mathrm{p}_{t,j}^G; s_{j,i})$.

$\ell = (\ell_j)_{j=1}^{J-1}$ as the function allowing us to compute the length of the segments of a pose p:

$$\ell_j : \mathrm{p} \mapsto \|\mathrm{p}_j - \mathrm{p}_{\tau(j)}\|_2 \,, \quad 0 < j \leq J-1 \,, \tag{11}$$

where $\tau : \{1, \ldots, J-1\} \to \{0, \ldots, J-1\}$ maps joint indices to the index of their parent joint:

$$\tau(i) = j < i, \quad \text{s.t. } A_{ij} = 1 \,. \tag{12}$$

From assumption 5, we know that for any pose p from the training distribution,

$$\forall j \,, \quad \ell_j(\mathrm{p}) = s_{j,\tau(j)} \,. \tag{13}$$

Given $D = \{(\mathrm{x}_i, \mathrm{p}_i)\}_{i=1}^N \sim \mathrm{P}(\mathrm{x}, \mathrm{p})$, some i.i.d. evaluation data, the MSE of a model f is defined as:

$$\mathrm{MSE}(f; N) = \frac{1}{N} \sum_{i=1}^{N} \|\mathrm{p}_i - f(\mathrm{x}_i)\|_2^2 \,, \tag{14}$$

and converges to

$$\mathrm{MSE}^*(f) = \mathbb{E}_{\mathrm{x},\mathrm{p}}\left[\|\mathrm{p} - f(\mathrm{x})\|_2^2\right] \tag{15}$$

as the dataset size N goes to infinity. We then define the oracle MSE minimizer as

$$f^* = \arg\min_f \mathrm{MSE}^*(f) \,. \tag{16}$$

The quantity in (15) is known in statistics as the expected L_2-risk and it is a well-known fact that its minimizer is the conditional expectation:

$$f^*(\mathrm{x}) = \mathbb{E}[\mathrm{p}|\mathrm{x} = \mathrm{x}] \,. \tag{17}$$

Thus, since ℓ_j^2 are strictly convex and $\mathrm{P}(\mathrm{p}|\mathrm{x})$ is non-degenerate according to assumption 6, we can conclude from Jensen's strict inequality that for all j,

$$\ell_j^2(f^*(\mathrm{x})) = \ell_j^2(\mathbb{E}[\mathrm{p}|\mathrm{x} = \mathrm{x}]) < \mathbb{E}[\ell_j^2(\mathrm{p})|\mathrm{x} = \mathrm{x}] = s_{j\tau(j)}^2 \,, \tag{18}$$

where the last equality arises from the fact that $\ell_j^2(\mathrm{p})$ is not random according to (13). Thus, given that $\ell_j > 0$ and $s_{j,\tau(j)} > 0$, we can say that $\ell_j(f^*(\mathrm{x})) < s_{j,\tau(j)}$ for all joints j. We conclude that the model f^* minimizing MSE^* predicts poses that violate assumption 5 and are inconsistent. ■ ■

As an immediate corollary of proposition 1, we may state the following result, which was empirically illustrated in many parts of our paper:

Corollary 8 *Given a fixed training distribution* $\mathrm{P}(\mathrm{x}, \mathrm{p})$ *respecting assumptions 4-6, for all 3D-HPE model* f *predicting consistent poses,* i.e. *, that respect assumption 5, there is an inconsistent model* f' *with lower mean-squared error.*

Proof Let $f' \in \arg\min_{\tilde{f}} \mathrm{MSE}^*(\tilde{f})$. According to proposition 1, f' is inconsistent. Suppose that the consistent model f is such that

$$\mathrm{MSE}^*(f) \leq \mathrm{MSE}^*(f')\,. \tag{19}$$

Since MSE^* reaches its minimum at f', we have $\mathrm{MSE}^*(f) = \mathrm{MSE}^*(f')$. Thus, $f \in \arg\min_{\tilde{f}} \mathrm{MSE}^*(\tilde{f})$, which means that f is also inconsistent according to proposition 1. That is impossible given that we assumed f to be consistent. We conclude that Eq. (19) is wrong and that

$$\mathrm{MSE}^*(f) > \mathrm{MSE}^*(f')\,. \tag{20}$$

■ ■

Note that propositions 1 and 8 assume the use of the MSE loss, which is the most widely used loss in 3D-HPE. We can however extend them to the case where MPJPE serves as optimization criteria under an additional technical assumption:

Corollary 9 *The predicted poses minimizing the mean-per-joint-position-error loss are inconsistent if the training poses distribution* $\mathrm{P}(\mathrm{x}, \mathrm{p})$ *verifies Asm. 4-6 and if the joint-wise residuals' norm standard deviation is small compared to the joint-wise loss:*

$$0 \leq j < J\,, \quad \frac{\sqrt{\mathbb{V}_{\mathrm{x,p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]}}{\mathbb{E}_{x,\mathrm{p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]} \simeq 0\,. \tag{21}$$

Proof From proposition 1 we know that the poses predicted by the minimizer f^* of

$$\mathrm{MSE}^*(f) = \mathbb{E}_{\mathrm{x,p}}\left[\|\mathrm{p} - f(\mathrm{x})\|_2^2\right] \tag{22}$$

are inconsistent. Let f_j be the component of f corresponding to the j^{th} joint. We define the j^{th} mean-per-joint-position-error component as:

$$\mathrm{MPJPE}_j^*(f) \triangleq \mathbb{E}_{\mathrm{x,p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]. \tag{23}$$

Under the small variance assumption, we have:

$$\frac{\mathbb{V}_{\mathrm{x,p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]}{\mathbb{E}_{x,\mathrm{p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]^2} \tag{24}$$

$$= \frac{\mathbb{E}_{\mathrm{x,p}}\left[\|\mathrm{p} - f(\mathrm{x})\|_2^2\right] - \mathbb{E}_{\mathrm{x,p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]^2}{\mathbb{E}_{x,\mathrm{p}}\left[\|\mathrm{p}_j - f_j(\mathrm{x})\|_2\right]^2} \tag{25}$$

$$= \frac{\mathrm{MSE}_j^*(f) - \mathrm{MPJPE}_j^*(f)^2}{\mathrm{MPJPE}_j^*(f)^2} \simeq 0\,, \tag{26}$$

so both criteria, MSE and MPJPE, are asymptotically equivalent and have the same minimizer f^*, which is inconsistent according to proposition 1. ■ ■

Appendix D. Further details on the 1D-to-2D case study

D.1. Implementation details

Datasets. We created a dataset of input-output pairs $\{(x_i, (x_i, y_i))\}_{i=1}^{N}$, divided into 1 000 training examples, 1 000 validation examples and 1 000 test examples. Since the 2D position of J_1 is fully determined by the angle θ between the segment (J_0, J_1) and the x-axis, the dataset is generated by first sampling θ from a von Mises mixture distribution, then converting it into Cartesian coordinates (x_i, y_i) to form the outputs, and finally projecting them into the x-axis to obtain the inputs.

Distribution scenarios. We considered three different distribution scenarios with different levels of difficulty:

1. **Easy scenario**: a unimodal distribution centered at $\theta = \frac{2\pi}{5}$, where the axis of maximum 2D variance is approximately parallel to the x-axis (Fig. 2-A).
2. **Difficult unimodal scenario**: a unimodal distribution centered at $\theta = 0$, where the axis of maximum 2D variance is perpendicular to the x-axis (Fig. 2-B).
3. **Difficult multimodal scenario**: a bimodal distribution, with modes at $\theta_1 = \frac{\pi}{3}$ and $\theta_2 = -\frac{\pi}{3}$ and mixture weights $w_1 = \frac{2}{3}$ and $w_2 = \frac{1}{3}$, *i.e.*, where the projection of modes onto the x-axis are close to each other (Fig. 2-C).

All von Mises components in all scenarios had concentrations equal to 20.

Architectures and training. All three models were based on a multi-layer perceptron (MLP) with 2 hidden layers of 32 neurons each, using `tanh` activation.

The constrained and unconstrained MLPs were trained using the mean-squared loss $\frac{1}{N}\sum_{i=1}^{N}((\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2)$. ManiPose was trained with the loss in Eq. (1), and had $K = 2$ heads. We trained all models with batches of 100 examples for a maximum of 50 epochs. We used the Adam optimizer (Kingma and Ba, 2014), with default hyperparameters and no weight decay. Learning rates were searched for each model and distribution independently over a small grid: $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$ (*cf.* selected values in Table 5). They were scheduled during training using a plateau strategy of factor 0.5, patience of 10 epochs and threshold of 10^{-4}.

Table 5: **Selected learning rates for 1D-to-2D synthetic experiment.**

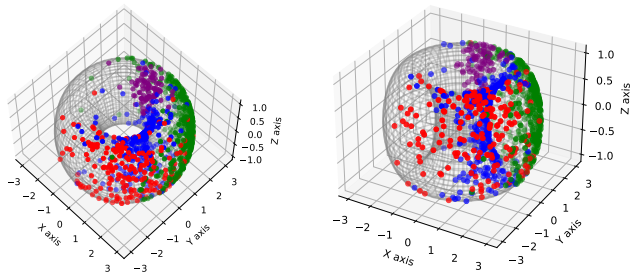
Distribution	A	B	C
Unconstr. MLP	10^{-3}	10^{-3}	10^{-2}
Constrained MLP	10^{-2}	10^{-4}	10^{-2}
ManiPose	10^{-2}	10^{-3}	10^{-2}

D.2. Extension to 2D-to-3D setup with more joints

We further extend the two-joint 1D-to-2D lifting experiment of Section 4.1 to 2D-to-3D with three joints, aiming at providing a scenario that is closer to real-world 3D-HPE, but that can still be fully dissected and visualized.

As in Section 4.1, we suppose that joint J_0 is at the origin at all times, that J_1 is connected to J_0 through a rigid segment of length s_1 and that J_2 is connected to J_1 through a second rigid segment of length $s_1 < s_0$. We further assume that both J_1 and J_2 are allowed to rotate around two axes orthogonal to each other. Thus, J_1 is constrained to lie on a circle $S^1(0, s_0)$, while J_2 lies on a torus $\mathcal{T}$ homeomorphic to $S^1(0, s_0) \otimes S^1(0, s_1)$. Without loss of generality, we set the radii $s_0 = 2$ and $s_1 = 1$ and assume them to be known.

Given that setup, we are interested in learning to predict the 3D pose $(J_1, J_2) = (x_1, y_1, z_1, x_2, y_2, z_2) \in \mathbb{R}^6$, given its 2D projection $(K_1, K_2) = (x_1, z_1, x_2, z_2) \in \mathbb{R}^4$. We create a dataset comprising $20,000$ training, $2,000$ validation, and $2,000$ test examples, sampled using an arbitrary von Mises mixture of poloidal and toroidal angles (θ, ϕ) in $\mathcal{T}$. We set the modes of such a mixture at $[(-\pi, 0), (0, \pi/4), (\frac{1}{2}, -\pi/4), (2\pi/3, \pi/2)]$, with concentrations of $[2, 4, 3, 10]$ and weights $[0.3, 0.4, 0.2, 0.1]$. Similarly to Fig. 2-C, that creates a difficult multimodal distribution, depicted in Fig. 8.



(a) (b)

Figure 8: **Visualisation of the von Mises mixture distribution on the torus T.** The different colors (blue, green, red, purple) represent the modes of the sampled points. We are only representing joint J_2 here for clarity.

We train and evaluate the same baselines as in Section 4.1 in that new scenario, using a similar setup (*cf.* Appendix D.1, Architectures and training). The corresponding Mean Per Segment Consistency Error (MPSCE) and Mean Per Joint Position Error (MPJPE) results are reported in Table 6.

Table 6: **Mean per joint prediction error (MPJPE) and mean per segment consistency error (MPSCE) in a 2D-to-3D scenario.** ManiPose reaches perfect MPSCE consistency without degrading MPJPE performance.

	MPJPE ↓	MPSCE ↓
Unconst. MLP	1.1468	0.2539
Constrained MLP	1.1593	**0.0000**
ManiPose	**1.1337**	**0.0000**

We see that the same observations as in Section 4.1 also apply here: although the unconstrained MLP yields competitive MPJPE results, its predictions are not consistently aligned with the manifold, as indicated by its poor MPSCE performance. Again, we show

here that ManiPose offers an effective balance between maintaining manifold consistency and achieving high joint-position-error performance.

Appendix E. Further ManiPose implementation details

E.1. Architectural details

Our architecture is backbone-agnostic, as shown on Fig. 1. Thus, in order to have a fair comparison, we decided to implement it using the most powerful architecture available, *i.e.*, MixSTE (Zhang et al., 2022).

In practice, the rotations module follows the MixSTE architecture with $d_l = 8$ spatio-temporal transformer blocks of dimension $d_m = 512$ and time receptive field of $T = 243$ frames for Human 3.6M experiments and $T = 43$ frames for MPI-INF-3DHP experiments. Contrary to MixSTE, that network outputs rotation embeddings of dimension 6 for each joint and frame, instead of Cartesian coordinates of dimension 3.

Concerning the segment module, it was also implemented with a smaller MixSTE backbone of depth $d_l = 2$ and dimension $d_m = 128$.

The ablation study presented in Table 4 shows that the increase in the number of parameters between MixSTE and ManiPose is negligible.

E.2. Pose decoding details

The pose decoding block from Fig. 1 is described in Section 2.1 and is based on Algorithms 1 and 2. The whole procedure is illustrated on Fig. 4.

Table 7: Joint-wise weights used in the Winner-takes-all loss Eq. (2) (as in Zhang et al. (2022)).

Joint	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Weight	1	1	2.5	2.5	1	2.5	2.5	1	1	1	1.5	1.5	4	4	1.5	4	4

Algorithm 1 6D rotation representation conversion (Zhou et al., 2019)

Require: Predicted 6D rotation representation $r \in \mathbb{R}^6$.

$x' \leftarrow [r_0, r_1, r_2]$,
$y' \leftarrow [r_3, r_4, r_5]$,
$x \leftarrow x'/\|x'\|_2$,
$z' \leftarrow x \wedge y'$,
$z \leftarrow z'/\|z'\|_2$,
$y \leftarrow z \wedge x$,
return $R = [x|y|z] \in \mathbb{R}^{3\times3}$.

E.3. Training details

Training tactics. In order to have a fair comparison with MixSTE (Zhang et al., 2022), we trained ManiPose using the same training tactics, such as pose flip augmentation both

Algorithm 2 Forward Kinematics (Murray et al., 2017; Li et al., 2021)

Require: Scaled reference pose $u' \in (\mathbb{R}^3)^J$, predicted rotation matrices $R_{t,j}$, $0 \leq j < J$.

$R'_{t,0} \leftarrow R_{t,0}$,

$\mathrm{p}_{t,0} \leftarrow \mathrm{u}'_0$, **for** $j = 1, \ldots, J-1$ **do**

end

$R'_{t,j} \leftarrow R_{t,j} R'_{t,\tau(j)}$, ▷ Compose relative rotations

$\mathrm{p}_{t,j} \leftarrow R'_{t,j}(\mathrm{u}'_j - \mathrm{u}'_{\tau(j)}) + \mathrm{p}_{t,\tau(j)}$,

return $\mathrm{p}_t = [\mathrm{p}_{t,j}]_{0 \leq j < J}$

at training and test time. Moreover, the training loss (1) was complemented with two additional terms described in Zhang et al. (2022):

1. a TCloss term, initially introduced in Hossain and Little (2018);
2. a velocity loss term, introduced in Pavllo et al. (2019).

We also weighted the Winner-takes-all MPJPE loss (2) as in Zhang et al. (2022) (*cf.* weights in Table 7). The score loss weight, β, was set to 0.1, while TCloss and velocity loss terms had respective weights of 0.5 and 2 (values from Zhang et al. (2022)).

Training settings. We trained our model for a maximum of 200 epochs with the Adam optimizer (Kingma and Ba, 2014), using default hyperparameters, a weight decay of 10^{-6} and an initial learning rate of 4×10^{-5}. The latter was reduced with a plateau scheduler of factor 0.5, patience of 11 epochs and threshold of 0.1 mm. Batches contained 3 sequences of $T = 243$ frames each for the Human 3.6M training, and 30 sequences of $T = 43$ frames for MPI-INF-3DHP.

Compute resources. Trainings were carried out on a single NVIDIA RTX 2000 GPU with around 11GB of memory. The training of the large model with 243 frames on Human 3.6M dataset took around 26 hours.

Dataset licences. Human 3.6M is a dataset released under a research-only custom license, and is available upon request at this URL: http://vision.imar.ro/human3.6m/description.php. MPI-INF-3DHP is released under non-commercial custom license and can be found at: https://vcai.mpi-inf.mpg.de/3dhp-dataset/.

E.4. Baselines evaluation.

All Human 3.6M evaluations of MPSSE and MPSCE listed in Tables 2 and 4 were performed using the official checkpoints of these methods and their corresponding official evaluation scripts. Concerning MPI-INF-3DHP evaluations from Table 3, checkpoints were not available (except for P-STMO). Thus, baseline models were retrained from scratch using the official MPI-INF-3DHP training scripts provided by the authors of each work, using hyperparameters reported in their corresponding papers. We checked that we were able to reproduce the reported MPJPE results.