MoManifold: Learning to Measure 3D Human Motion via Decoupled Joint Acceleration Manifolds

Technical Appendix

Outline

Here we provide details, extended experiments and ablation studies omitted from the main paper for brevity. App. A provides implementation details, App. B gives the experimental evaluation details, App. C presents more experiments of our motion prior, App. D contains our ablation studies and App. E provides some extended discussions. We encourage the reader to view the supplementary video for more qualitative results.

A Implementation Details

A.1 Weighted Design

Here, we introduce the weighted design of Eq. (3) in the main paper, where w_i is determined by the summation of bone lengths from joint i to the root joint along the kinematic structure of the SMPL body model. For joint i, the summation of bone lengths l_i is,

$$l_i = \sum b, \tag{1}$$

where b is the bone length. Thus, through experimental exploration, w_i is defined as:

$$w_i = \frac{4l_i^2}{4l_i^2 + 1}. \tag{2}$$

This design ensures that joints with intenser movements contribute proportionally more to the unsigned distance field of motion segment $\mathbf{m}$.

A.2 Data Preparation

Training Data. The training data is divided into two categories: plausible motion data and noisy motion data. We use the train split of AMASS dataset [15] as the plausible motion data, *i.e.*, the zero level of plausible acceleration vectors manifolds. We downsample AMASS to 25Hz or 24Hz because it records human motion at 100Hz or 120Hz. This will ensure that the temporal gap of consecutive frames between the two frequency motion data is closest and it can be easily generalized to higher frequencies *e.g.*, 30Hz. Then, we randomly sample motion segments of fixed lengths to model the manifolds.

For the noisy motion data, which lies outside the manifold, we utilize artificially noised motion data and the results from a representative SMPL-based human pose estimator VIBE [10]. We apply the noise from a uniform distribution, rather than Gaussian noise, to create artificially noised motion data of AMASS training set. Because it will produce a more diverse and wider distribution of noisy motion. Specifically, for a motion segment of length y, we randomly select x $(x \leq y)$ frames for adding noise, where x is also randomly generated. Furthermore, after employing manually noised motion data for training, we performed fine-tuning using the results from the human pose estimator VIBE on videos of MPI-INF-3DHP dataset [16]. Due to self-occlusion and partial observations, the estimates output by existing estimators encompass a substantial amount of noisy motion that is hard to be replicated through artificial noise. Additionally, such noisy motion is closer to the manifold, which will help learn a more refined manifold surface. Notably, we do not use any ground truth annotations from the MPI-INF-3DHP dataset.

We employ KNN algorithm [2] to compute the ground truth distance values of acceleration vectors outside the manifold. We implement KNN using FAISS [7]. Specifically, for an acceleration vector, we calculate the top-k nearest distances to the zero level and then compute the average distance as the ground truth distance. In our setup, we use $k = 5$.

Evaluation Data. For the motion denoising experiments in Sec. 4.2, we utilize two real world mocap data HPS [5] and the test split of AMASS [15]. HPS records human motion at 30Hz, thus we do not perform downsampling and directly conduct the evaluation on the motion of 30Hz. However, for the AMASS dataset, which records human motion at 100Hz or 120Hz, we downsample it to 25Hz or 24Hz for the evaluation. For HPS dataset, we randomly sampled 150 motion sequences for each setup. And in the experiments of AMASS dataset, we randomly selected 100 motion sequences for 60 frames or 120 frames. However, for the 240 frames of AMASS, due to downsampling requirements, we could only randomly sample 71 motion sequences for evaluation. Then, following Pose-NDF [20], we introduce random noise to each frame to create noisy observations.

In the fitting to partial experiments of Sec. 4.3, we use the test split of AMASS for evaluation, which is also downsampled to 25Hz or 24Hz. We also randomly selected 100 motion sequences for 60 frames or 120 frames. To simulate occlusion, we randomly select one-third of the frames within a motion sequence and set the rotations of corresponding occluded joints to zero. Besides, during optimization, when calculating the observation alignment term *i.e.*, Eq. (12) in main paper, the occluded joints of occluded frames are excluded.

A.3 Optimization Details

Since our motion prior is built upon motion segments, for an entire motion sequence, we initially split it into distinct motion segments by employing a sliding window with the window size equal to the length of our prior and the stride of 1. This will avoid boundary effects and make any motion segment comply with human motion dynamics. Subsequently, we calculate the distance of each motion segment to the plausible motion manifold, and then utilize the average distance of these motion segments to guide the optimization process. For the experiments of motion denoising and fitting to partial observations, the optimization variable in Adam [9] is the entire motion sequence. Specially, for post-optimization of human pose estimators and motion in-betweening, as we only use our motion prior without any other optimization objectives, we optimize each motion segment individually, recording multiple results of each frame to obtain the final optimized poses with a weighted average strategy similar to SmoothNet. It will increase the receptive field of each frame during optimization.

B Experimental Evaluation Details

B.1 Datasets

We evaluate our motion prior on five datasets including AMASS [15], HPS [5], 3DPW [22], AIST++ [13] and LAFAN1 [6].

AMASS is a large motion capture database containing diverse motion and body shapes on the SMPL body model. We sub-sample the dataset to 25Hz or 24Hz and use the recommended training split to train the unsigned distance fields. For the evaluation data, we also perform the same downsampling on the test split of AMASS.

HPS is a method to recover the full 3D pose of a human registered with a 3D scan of the surrounding environment using wearable sensors. And with this method, HPS recorded several large 3D scenes (300-1000 sq.m) consisting of 7 subjects and more than 3 hours of diverse motion.

3DPW is a challenging in-the-wild dataset consisting of 60 videos, which are captured by a phone at 30 FPS. Moreover, IMU sensors are utilized to obtain the near ground-truth SMPL parameters, *i.e.*, pose and shape.

AIST++ is a challenging dataset that comes from the AIST Dance Video DB [21]. It contains 1,408 sequences of 3D human dance motion, represented as joint rotations along with root trajectories.

LAFAN1 is a high-quality public motion capture dataset. It contains 15 actions performed by 5 actors such as walking, dancing, fighting, jumping, with 496,672 frames captured in a production-grade motion capture system at 30Hz. We adopt the same test set in [18], which contains 2,232 clips sampled with a window of 65, offset by 40 frames on Subject 5. Although this dataset is not based on SMPL, its human skeleton definition is completely consistent with SMPL and joint rotations are provided, so the poses of this dataset can be converted into SMPL poses. In addition, since the rest-pose of this dataset is not T-Pose, the relative rotations of the joints in the dataset cannot be directly converted to those of SMPL, so we first converted the dataset so that all joint rotations are all relative to T-Pose.

B.2 Evaluation Metrics

For the evaluation, five standard metrics are used, including MPJPE, PA-MPJPE, PVE, and Accel.

MPJPE (Mean Per Joint Position Error) is calculated as the mean of the Euclidean distance between the ground-truth and the estimated 3D joint positions after aligning the pelvis joint on the ground truth location. MPJPE comprehensively evaluates the predicted poses and shapes, including the global orientations.

PA-MPJPE (Procrustes-Aligned Mean Per Joint Position Error) performs Procrustes alignment before computing MPJPE, which mainly measures the articulated poses, eliminating the differences in scale and global orientation.

PVE (Mean Per Vertex Position Error) is calculated as the mean of the Euclidean distance between the ground truth and the estimated 3D human mesh vertices (output by the SMPL model).

Accel (Mean Per Joint Acceleration Error) is measured as the mean difference between the ground-truth and the estimated 3D acceleration for every joint. It is used to express the smoothness and temporal coherence of 3D human motion as well as the similarity to ground-truth motion.

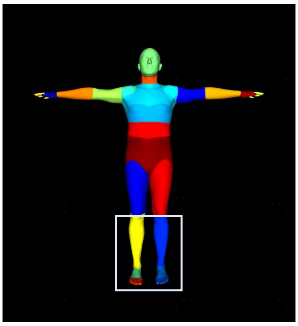


Figure S1: **SMPL Body Segmentation.** The white box contains the segmentation of legs, feet and toe-bases.

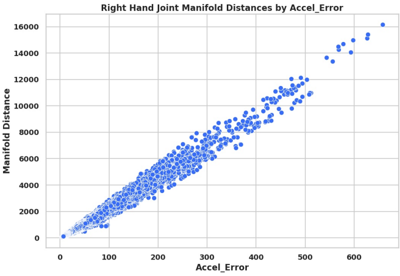


Figure S2: **SMPL Body Segmentation.** The white box contains the segmentation of legs, feet and toe-bases.

NPSS (The Normalized Power Spectrum Similarity) proposed by [3], evaluates angular differences between predicted motion and ground truth on the frequency domain. NPSS measures similarity of motion patterns, which reportedly correlates better with human perception of quality.

B.3 PVE of Legs and Feet

In Table 5 of the main paper, we employ the PVE (Per Vertex Error) of legs and feet to numerically demonstrate that our method avoids resulting in footskate when smoothing the motion, compared with SmoothNet [23]. As shown in Figure S1, we segment the human body mesh into different parts through the indices of mesh vertices provided by [14] and then compute the PVE for the vertices belonging to the legs, feet and toe-bases (in mm).

B.4 Optimization Space of Rotation

For fair comparison, in Sec. 4.2 and Sec. 4.3, we optimize the human poses in the axis-angle space same with Pose-NDF [20], and in Sec. 4.4, we adopt the space of 6D rotation representation [24] following SmoothNet [23]. Moreover, we have observed that optimizing human poses in the 6D space is more stable and leads to better convergence in some cases compared to the axis-angle space. Therefore, in Sec. 4.5 and Sec. C.5, we optimize the human poses in the 6D space.

C Extended Experiments

For dynamic motion and better qualitative comparison, we recommend viewing our supplementary video.

C.1 Correlation Analysis

In this section, we will present the intuitive visualization of the positive linear correlations between the manifold distances and acceleration error across joints. The linear correlation of the right hand joint are visualized in Figure S2. Moreover, Figure S3 shows the linear correlations of the other joints. The two joints on the spine and the head joint are missing here because there are no corresponding joints in the GT skeleton of 3DPW, so acceleration error cannot be obtained.

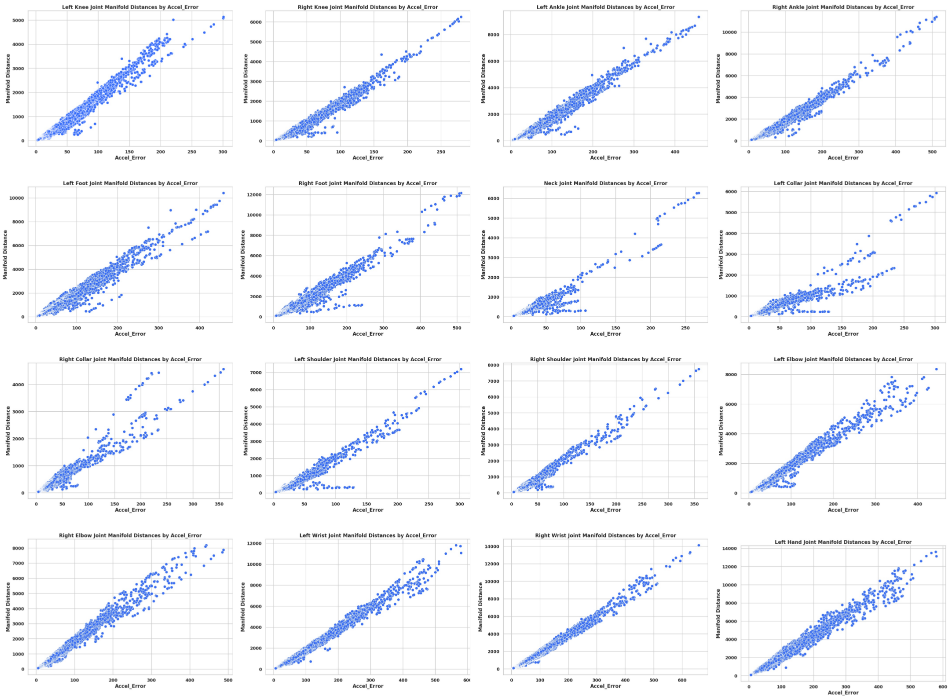


Figure S3: **Scatter Plots of Other Joints.** Each blue point represents a motion segment.

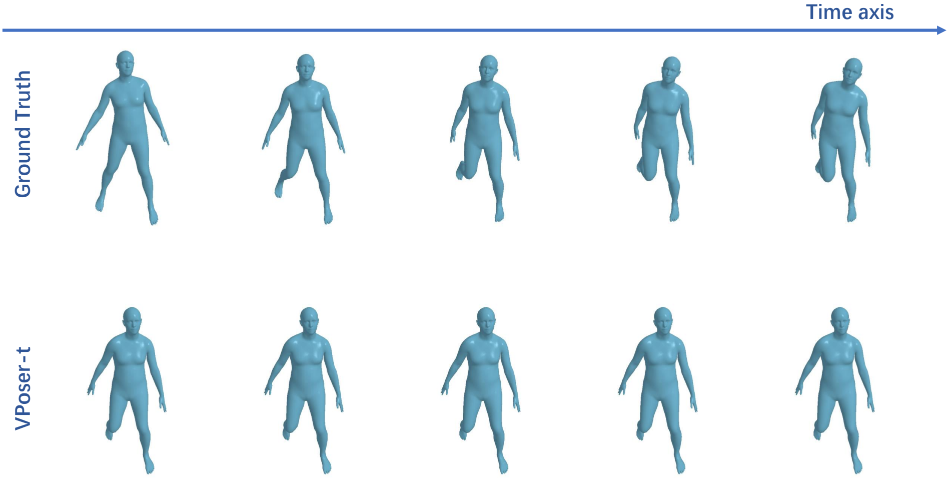


Figure S4: **VPoser-t Denoising Results.**

C.2 Motion Fitting from 3D Observations

VPoser-t [17] embeds human poses into a biased Gaussian space of VAE-based representations and optimizes poses within the latent space, resulting in average poses. When these average poses are assembled into motion, the resulting sequences appear stiff and mechanical as shown in Figure S4.

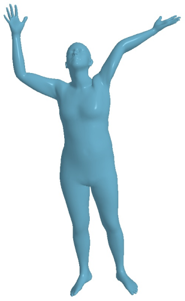


Figure S5: **HuMoR Accumulation of Errors.**

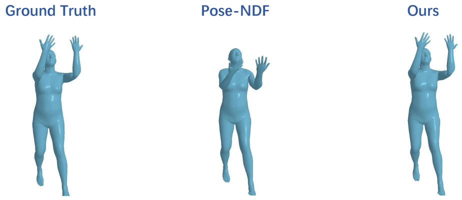


Figure S6: **Motion Range Comparison.** Due to constraints imposed by the traditional temporal regularization term, Pose-NDF struggles to achieve the correct height for arm elevation. In contrast, our method could preserve a more realistic range of motion.

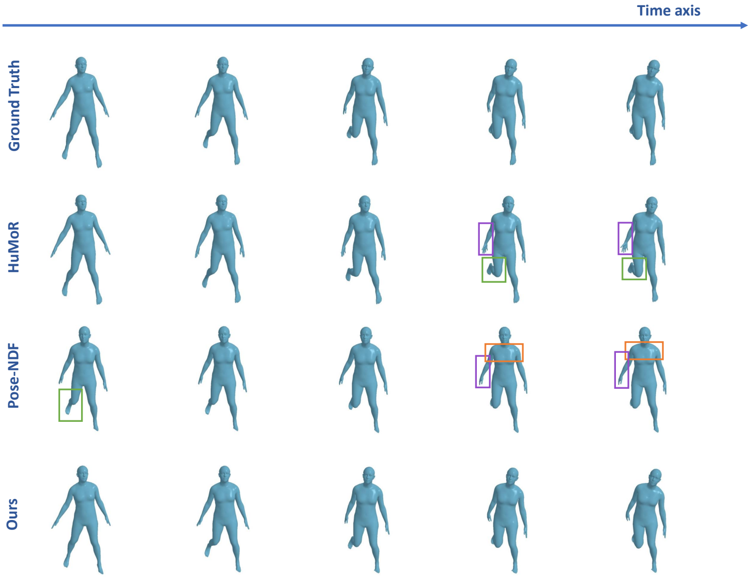


Figure S7: **Denoising Comparison.** Body parts that are significantly different from the ground truth are marked in colored boxes. The results of VPoser-t are same with Figure S4. For uniform motion of Pose-NDF, the legs begin to retract in the first frame, whereas at that time, the human should stand on the ground. Besides, the right arm and shoulders in the last two frames are obviously different from the ground truth. Since this is the beginning of the motion, there is no accumulation of errors for HuMoR. And our results are the closest to the ground truth.

HuMoR [19] could also recover realistic motion in some cases, but due to modeling of transitions between only two consecutive frames, there might be an accumulation of errors leading to extreme unrealistic poses (as shown in Figure S5) in the final few frames of the motion, which has also been demonstrated in [20].

Pose-NDF [20] employs a traditional temporal regularization term to smooth motion, but this tends to cause the uniform motion. Because the optimization direction of such temporal terms aims to minimize the frame-to-frame differences, effectively freezing the motion. Hence, the motion generated by Pose-NDF exhibits minimal variation in velocity, which will result in a lack of dynamism, particularly in actions that involve distinct changes in speed,

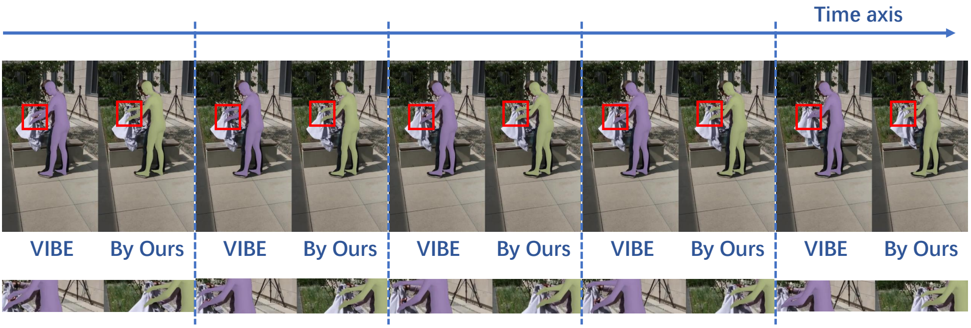


Figure S8: **Post-optimization for Human Pose Estimator VIBE.** This figure displays a motion sequence of five consecutive frames. The bottom row shows the enlarged images of the arms. We can see that VIBE produced a sudden jitter of the right arm, while through our optimization, we can mitigate the jitter issues.

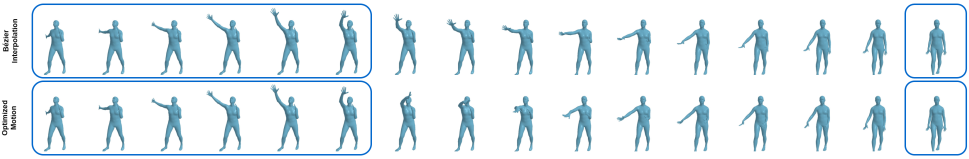


Figure S9: **Qualitative Comparison with Bézier Interpolation.** Frames 0 to 5 and frame 15 in the blue boxes are the conditional poses.

such as pushing movements. Furthermore, the range of motion will also be restricted, as depicted in Figure S6.

In Figure S7, we present the initial five frames of the side hopping motion, and the results of our method are closest to ground truth since we can well preserve the human motion dynamics. We suggest watching our supplementary video for more qualitative results.

C.3 Mitigating Jitters for SMPL-based Pose Estimators

MoManifold learns an unsigned distance field of plausible motion and explicitly quantifies human motion dynamics into a score (*i.e.*, distance) which can guide the optimization process. Therefore, our motion prior can be utilized to mitigate jitter issues produced by existing human pose estimators because the motion with jitter movements must be outside the manifold of plausible motion and has a large distance. In Figure S8, we present a qualitative comparison with a representative human pose estimator VIBE. For more qualitative results, please refer to our supplementary video. Besides, as shown in Table S1, the estimation performance often degrades when applying traditional filters (such as one euro) which has been proven in [23].

C.4 Motion In-betweening Refinement

Moreover, we also evaluate our method with first-order Bézier (linear) interpolation, commonly used in animation software. Specifically, we select frames 0 to 5 and frame 15 as conditional poses which are randomly sampled from AMASS and adopt Bézier interpolation for initial in-betweening, and then we further optimize it with our motion prior. The results are shown in Figure S9. We can see that our method captures human motion dynamics better by

Method	3DPW			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
SPIN [[redacted]]	99.29	61.71	113.32	34.95
SPIN w/ one euro	99.53	62.24	113.55	14.23
SPIN w/ S [[redacted]]	97.81	61.19	111.5	**7.4**
SPIN w/ only proposed	**97.28**	**60.79**	**111.4**	8.55
EFT [[redacted]]	91.6	55.33	110.17	33.38
EFT w/ one euro	91.82	55.65	110.46	14.17
EFT w/ S [[redacted]]	89.57	54.40	**107.66**	**7.89**
EFT w/ only proposed	**89.48**	**53.91**	107.94	9.05
PARE [[redacted]]	79.93	48.74	94.07	26.45
PARE w/ one euro	80.46	49.32	94.81	10.52
PARE w/ S [[redacted]]	78.68	48.47	**92.5**	**6.31**
PARE w/ only proposed	**78.61**	**47.86**	92.72	7.75
VIBE* [[redacted]]	84.28	54.93	99.10	23.59
VIBE* w/ one euro	85.89	56.49	100.80	10.87
VIBE* w/ S [[redacted]]	83.46	54.83	98.04	**7.42**
VIBE* w/ only proposed	**83.14**	**54.29**	**97.87**	8.12
TCMR* [[redacted]]	88.47	55.70	103.22	7.13
TCMR* w/ one euro	90.18	57.41	104.97	6.74
TCMR* w/ S [[redacted]]	88.69	56.61	103.40	**6.48**
TCMR* w/ only proposed	**88.28**	**55.69**	**103.02**	6.72

Table S1: **Mitigating Jitters on 3DPW Dataset.** "w/ one euro" refers to using the traditional one euro filter for refinement. "w/ S" indicates refinement using SmoothNet. "*" denotes spatio-temporal backbones.

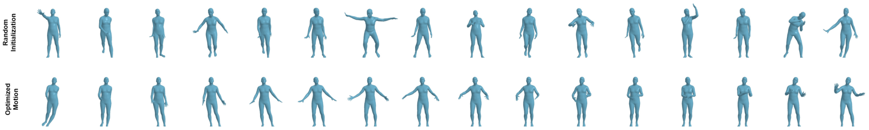


Figure S10: **Motion Generation.** The first row is the randomly initialized chaotic motion. The second row is the realistic motion we generated, which is the action of closing and subsequently spreading the hands.

guiding the optimization with manifold distances. Bézier interpolation only considers two key frames, while our motion prior takes into account the overall motion trend, so that the right arm still maintains a certain swinging motion before putting it down.

C.5 Motion Generation

Beyond enhancing the motions produced by existing methods, our approach even has a certain capability of motion generation by converting chaotic sequences into plausible human motions. We begin by randomly selecting 16 varied poses from the AMASS dataset, forming an initial erratic sequence. We then exclusively apply our motion prior, as defined in Eq. (8) of the main paper, to this disordered starting point. As illustrated in Figure S10, the generated motion is seamless and natural.

D Ablation Studies

D.1 Optimal Motion Segment Length

In this section, we perform the ablation study on the experiment of mitigating jitters for human pose estimators. We aim to find the optimal motion segment length. The length of the motion segment L determines the capacity of temporal information. Longer motion segments contain more temporal information, but also raise the modeling difficulty and manifold complexity. We demonstrate the effects on different lengths from 5 to 32 frames in Table S2. We

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
VIBE	84.28	54.93	99.10	23.59
VIBE w/ M-5	83.35	54.50	98.16	9.17
VIBE w/ M-8	83.17	54.34	97.92	8.30
VIBE w/ M-16	**83.15**	**54.30**	**97.88**	**8.18**
VIBE w/ M-32	83.32	54.48	98.11	8.44

Table S2: **Impact of Motion Segment Length.** We employ MoManifold to optimize the results of VIBE on 3DPW. "M-n" refers to using an n-frame motion segment to model the acceleration manifolds.

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel ↓
VIBE	84.28	54.93	99.10	23.59
VIBE w/ T	84.59	56.33	99.40	**7.50**
VIBE w/ M-16	83.15	54.30	97.88	8.18
VIBE w/ F-16	**83.07**	**54.28**	**97.80**	8.01

Table S3: **Impact of Different Temporal Terms.** Through fusion, MoManifold can achieve better performance. "w/ T" denotes with Traditional and "w/ F-16" means that we integrate the traditional term with M-16.

Data	**Noisy HPS**		**Noisy AMASS**	
# frames	60	120	60	120
VPoser-t [[redacted]]	3.05	4.43	5.83	6.55
HuMoR [[redacted]]	6.08	12.67	10.28	12.63
Pose-NDF [[redacted]]	1.17	1.30	5.03	5.39
Only proposed	**0.97**	**0.98**	**1.56**	**1.59**

Table S4: **Motion Denoising**. We compare PVE in cm. "Only proposed" refers to only using our motion prior to regularize the motion without integrating with the traditional temporal term.

Data	**Occ. Leg**		**Occ. Arm**		**Occ. Shoulder**	
			+Hand		**+Upper Arm**	
# frames	60	120	60	120	60	120
VPoser-t [[redacted]]	8.69	10.77	8.79	10.70	8.74	10.20
HuMoR [[redacted]]	9.52	12.70	9.39	13.82	9.02	12.14
Pose-NDF [[redacted]]	8.50	9.40	8.66	9.43	8.73	9.47
Only proposed	**5.09**	**5.33**	**5.06**	**5.26**	**5.19**	**5.32**

Table S5: **Fitting to Partial Data.** We compare PVE (in cm) on test set of AMASS. Even without the integration, our results are still better than other methods in all cases.

chose 5 as the minimum motion segment length because the acceleration vector empirically should be at least 3 frames. Table S2 shows that as the motion segment length increases, all four metrics first decrease and then begin to increase. When the motion segment length L is 16, we can obtain the best performance.

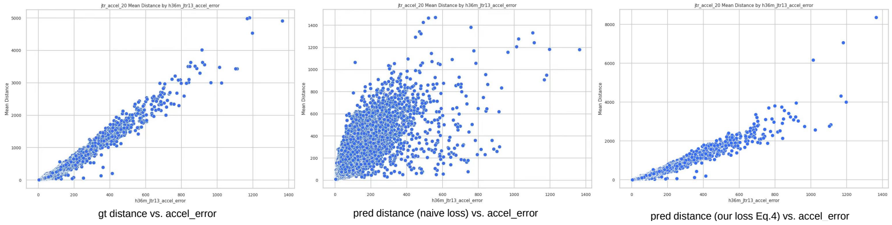
D.2 Impact of Different Temporal Terms

In this section, we explore the influence of different temporal terms on the experiment of mitigating jitters for human pose estimators. In the optimization-based tasks, various similar temporal regularization terms (*e.g.*, the sum of joint differences or mesh vertex differences between consecutive frames) are applied to smooth motion. Table S3 shows that, naively applying the traditional temporal regularization term Eq. (7) to optimize the pose estimator's results can indeed reduce acceleration error and mitigate jitter issues. However, it will lower human pose recognition accuracy, as indicated by MPJPE, PA-MPJPE, and PVE metrics. In contrast, by only utilizing Eq. (8), our method can not only mitigate jitter issues and smooth motion but also further enhance the pose recognition accuracy. Furthermore, we can see that the full optimization function, *i.e.*, an integration of both MoManifold and a traditional temporal regularization term, will further improve the performance, because it can help jump out of local optima during the optimization process.

D.3 Only Utilizing Proposed Prior

For the experiments of Sec. 4.2, Sec. 4.3 and Sec. 4.4 in the main paper, we used Eq. (9) to regularize motion, which integrates our motion prior with a traditional temporal regularization term. Here, we only use the proposed prior (*i.e.*, Eq. (8) in the main paper) in the experiments to demonstrate that even without the integration, we can still outperform the existing SOTAs as shown in Table S1, Table S4 and Table S5.

For the experiments of Sec. 4.5 and Sec. C.5, we exclusively apply our motion prior (without the traditional temporal term) as stated in the main paper.



gt distance vs. accel_error

pred distance (naive loss) vs. accel_error

pred distance (our loss Eq.4) vs. accel_error

Figure S11: **Ablation on Eq. (4).** This is correlation analysis of joint 20, same with Sec. 4.1. The left one corresponds to gt distances, the middle one corresponds to predicted distances with naive loss, and the right one corresponds to predicted distances with our loss.

D.4 Ablation on Losses

In this section, we conduct the ablation on the loss of Eq. (4). In Eq. (4), the logarithmic function changes first steeply and then gently, reducing large enough distances to similar values. This makes the neural network easier to learn, as it will pay more attention to points close to the manifold and will not be affected by points far away. In other words, Eq. (4) performs non-linear scaling for small and large distances. Figure S11 presents an intuitive comparison, proving that our loss function enables more accurate regression learning (the right one), whereas using a naive loss leads to inaccurate distance predictions (the middle one), thereby making it impossible to reflect the positive correlation with acceleration errors (the left one). For the loss of Eq. (5), [4] has demonstrated it would encourage a smoother distance field with unit-norm gradient outside the manifold.

E Discussions

E.1 About Joints Decoupling

At first, we tried to treat the human body as a whole and used various architectures, including transformers, to model the manifold, but it is hard to learn to map such high-dimensional input to a continuous distance value since extremely large data is required, which is impractical. Therefore, we proposed to decouple the joints, reducing the input dimension from 1008 to 42 (taking 16 frames as an example). This makes the data in the low-dimensional space dense enough to capture the data distribution.

Despite the decoupling, the joints maintain an inherent correlation through the SMPL model topology and thus reflect human dynamics as a whole. Indeed, this may make it hard to capture the kinematic relationships between joints on different branches, such as left leg and right arm. However, this will not cause pose errors when optimizing all joints, since we can always get the correct human body structure via SMPL model.

E.2 Joints J0-J3 are excluded

J0 is pelvis, the root joint, which corresponds to the position in the world coordinate system. J1 is left hip, J2 is right hip and J3 is spine1. Like previous methods, we set J0 fixed to better capture the changes of human poses in the local coordinate system. So J0 is static. J1-J3 are right next to J0 in the articulated skeleton and therefore have very little movement and very small acceleration, which makes it hard and meaningless to learn distance mapping.

References

[1] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021.

[2] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[3] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019.

[4] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020.

[5] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[6] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. 39(4), 2020.

[7] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[8] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020.

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

[10] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021.

[12] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[13] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[15] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[16] 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. doi: 10.1109/3dv.2017.00064. URL http://gvv.mpi-inf.mpg.de/3dhp_dataset.

[17] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[18] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[19] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[20] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022.

[21] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, November 2019.

[22] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[23] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022.

[24] 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.