Contact-aware Human Motion Forecasting —Supplementary Material—

Wei Mao[1], **Miaomiao Liu**[1], **Richard Hartley**[1], **Mathieu Salzmann**[2,3]
[1]Australian National University; [2]CVLab, EPFL; [3]ClearSpace, Switzerland
{wei.mao, miaomiao.liu, richard.hartley}@anu.edu.au, mathieu.salzmann@epfl.ch

1 Datasets

1.1 License

Both GTA-IM [1] and PROX [2] datasets are for non-commercial purpose only. For the details about their individual licenses, please follow the links below,

- GTA-IM: `https://github.com/ZheC/GTA-IM-Dataset/blob/master/LICENSE`
- PROX: `https://prox.is.tue.mpg.de/license.html`

1.2 Temporal Refinement of PROX

As described in the main script, the original PROX [2] dataset only provides jittery human motions which is ill-suited to our task. To obtain more smooth and realistic motions, we further process the provided dataset by applying the temporal smoothness constraints via an optimisation approach. Specifically, for every P consecutive frames, we would like to refine the original global rotation $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_P] \in \mathbb{R}^{P\times 6}$, global translation $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_P] \in \mathbb{R}^{P\times 3}$ and SMPL-X [6] pose parameters $\Theta = [\theta_1, \theta_2, \cdots \theta_P] \in \mathbb{R}^{P\times 561}$.

We first extract the point cloud of human at every frame $\{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_P\}$ from the ground-truth depth maps where $\mathbf{o}_p \in \mathbb{R}^{N_p\times 3}$ is the point cloud at p-th frame. Note that, since some parts of the human are often occluded by the scene given the monocular RGB-D videos of PROX, we cannot obtain high quality SMPL-X parameters with only those point clouds. We then design a two-stage temporal optimization process which first refines the global orientations ($\mathbf{R}$ and $\mathbf{T}$) and then the pose parameters (Θ).

Given the SMPL-X pose parameters θ_p at p-th frame and the shape parameters $\beta \in \mathbb{R}^{10}$, one can obtain the mesh vertices of the human via the SMPL-X model as,

$$\mathbf{v}_p = \mathcal{S}(\theta_p, \beta) \ , \tag{1}$$

where $\mathbf{v}_p \in \mathbb{R}^{N_\text{v}\times 3}$ is the 3D coordinate of N_v mesh vertices and $\mathcal{S}$ represents the SMPL-X model.

To refine the global orientations, we define a point cloud objective function which is the Chamfer Distance between the ground-truth human point cloud and the human vertices.

$$E_\text{pcd} = \frac{1}{P}\sum_{p=1}^{P} \mathcal{C}(\mathbf{o}_p, \mathbf{v}_p\tilde{\mathbf{r}}_p^T + \mathbf{t}_p) \ , \tag{2}$$

where $\mathcal{C}$ represents the Chamfer Distance of two set of points and $\tilde{\mathbf{r}}_p \in \mathbb{R}^{3\times 3}$ is the rotation matrix computed from its 6-D representation $\mathbf{r}_p$.

[1]Note that, following [7], we use the 6-D representation of the global rotation which is originally proposed in [8]. For SMPL-X pose, we use 32 latent representation of the body pose from the pretrained VPoser [6] and 24 PCA coefficients of the hand pose.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We also adopt a smoothness prior which is the speed and acceleration of the global rotation and translation.

$$E_{\text{smooth}} = \frac{1}{P-1}\sum_{p=2}^{P} \|\Delta \mathbf{r}_p\|_2^2 + \|\Delta \mathbf{t}_p\|_2^2 + \frac{1}{P-2}\sum_{p=3}^{P} \|\Delta^2 \mathbf{r}_p\|_2^2 + \|\Delta^2 \mathbf{t}_p\|_2^2 \,, \tag{3}$$

where $\Delta \mathbf{r}_p = \mathbf{r}_p - \mathbf{r}_{p-1}$, $\Delta \mathbf{t}_p = \mathbf{t}_p - \mathbf{t}_{p-1}$ and $\Delta^2 \mathbf{r}_p = \Delta \mathbf{r}_p - \Delta \mathbf{r}_{p-1}$, $\Delta^2 \mathbf{t}_p = \Delta \mathbf{t}_p - \Delta \mathbf{t}_{p-1}$.

The optimization of global orientations can then be expressed as

$$\mathbf{R}^*, \mathbf{T}^* = \arg\min_{\mathbf{R},\mathbf{T}} E_{\text{pcd}} + 0.1 E_{\text{smooth}} \tag{4}$$

At the second stage, we would like to optimize the pose parameters. We reuse the point cloud objective computed with the optimal global orientations.

$$E_{\text{pcd}} = \frac{1}{P}\sum_{p=1}^{P} \mathcal{C}(\mathbf{o}_p, \mathbf{v}_p \tilde{\mathbf{r}}_p^{*T} + \mathbf{t}_p^*) \,, \tag{5}$$

we also define a similar smoothness prior with Θ as

$$E_{\text{smooth}} = \frac{1}{P-1}\sum_{p=2}^{P} \|\Delta \theta_p\|_2^2 + \|\Delta^2 \theta_p\|_2^2 \,, \tag{6}$$

where $\Delta \theta_p = \theta_p - \theta_{p-1}$ and $\Delta^2 \theta_p = \Delta \theta_p - \Delta \theta_{p-1}$.

The optimization of the poses is then

$$\Theta^* = \arg\min_{\Theta} E_{\text{pcd}} + 0.1 E_{\text{smooth}} \tag{7}$$

2 Baseline

2.1 Changes of SLT

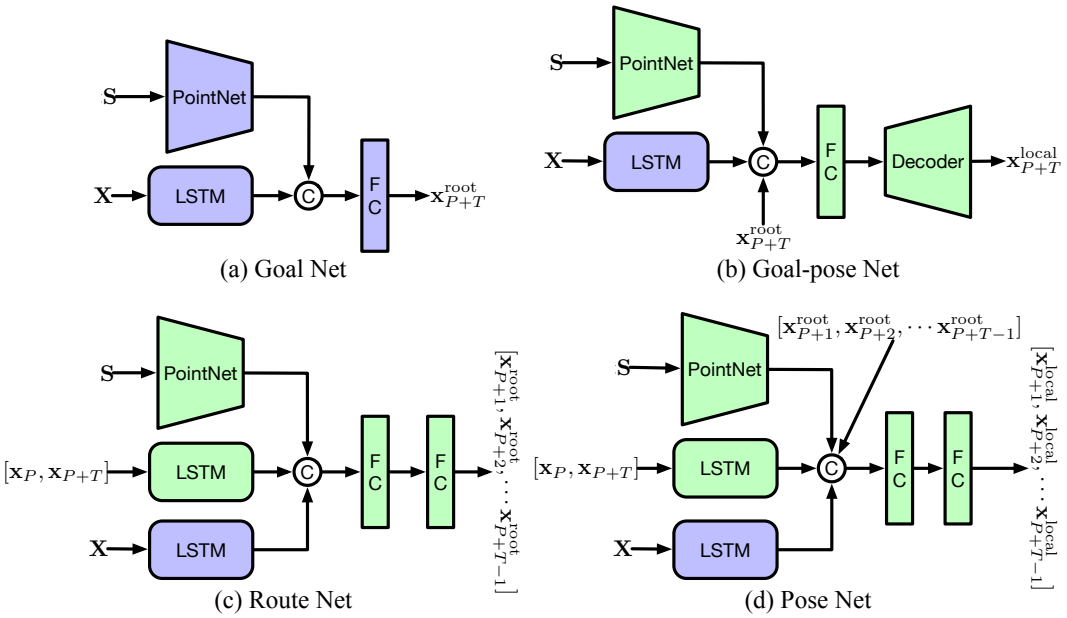
The official implementation of SLT [7] is for the task of human motion synthesis that is not applicable directly to our task. It is because 1) the generation process is not conditioned on any historical human motions; 2) in SLT [7], the model takes as input the ground-truth 3D position of the human at the goal frame; 3) SLT [7] relies on a VAE to synthesize human poses at the goal frame given the 3D position mentioned above. To achieve a fair comparison, we mainly made 3 changes on their networks to account for these settings mentioned above (shown in Fig. 1)

- We adapted the networks of SLT [7] to also take as input the embedding of the historical human motion. (Fig. 1 (a-d))
- We designed a new module to first predict the position of the human at the goal frame as no 3D location of the human is available in any future frame for the motion prediction task.(Fig. 1 (a))
- In this paper, we focus on deterministic human motion prediction, so that instead of using the VAE, we only use its decoder to predict the human poses at goal frame given the predicted 3D position. (Fig. 1 (b))

3 Implementation Details

To encode the past human poses, we use a 1-layer GRU with a hidden dimension of 128 for both contact prediction network and motion forecasting network. In our contact prediction net, we use the same backbone network of PVCNN architecture [4] with 2 modifications:

- We modified the decoder of PVCNN to also take as input the embedding of past human poses.



(a) Goal Net (b) Goal-pose Net

(c) Route Net (d) Pose Net

Figure 1: **Modified pipeline of SLT [7].** We show our changes in blue and the original module of SLT [7] in green. $\mathbf{S} \in \mathbb{R}^{N\times 3}$ is the 3D point cloud of the scene. $\mathbf{X} \in \mathbf{R}^{P\times J\times 3}$ is the past P human poses. (a) We added one network i.e., the Goal Net which aims to predict 3D translation of the last frame of the motion sequence ($\mathbf{x}^{\text{root}}_{P+T} \in \mathbb{R}^3$). (b) Given the scene $\mathbf{S}$, past human motion $\mathbf{X}$ and the global translation at last frame $\mathbf{x}^{\text{root}}_{P+T}$, instead of using a generative model i.e., VAE, we used the decoder of it to predict the local pose at that frame $\mathbf{x}^{\text{local}}_{P+T}$. (c)-(d) For the route net and pose net, we adapted the original models to also take an embedding of past motion as input.

- The output shape of our PVCNN is $N \times JL$, where N, J, L is the number of scene points, human joints and DCT coefficients, respectively.

In our motion forecasting network, we use a 6-layer MLP with 128 hidden units to predict the global translations. We also leverage the DCT representation with padding strategy. Specifically, the input to the MLP is the DCT coefficients of the padded historical translations. It aims to predict the real DCT coefficients. For both datasets, we retain the first 60 DCT coefficients. We then use a GRU with a hidden dimension of 128 to forecast the future poses. The GRU will take the embedding of past poses, the predicted global translations and the contact points as input.

4 Qualitative Results

Recall that, in the main context we only compare our results to those of DMGNN [3] and SLT [7]. Here, we provide a complete comparison to all baselines in Fig 2. In both datasets, the LTD [5] tends to produce ghost motions due to its lack of global and local motion constraints. Results are best viewed in videos, so that we provide more qualitative results in the supplementary video.

References

[1] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, pages 387–404. Springer, 2020.

[2] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019.

[3] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *CVPR*, pages 214–223, 2020.

[4] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *NeurIPS*, 32, 2019.

[5] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9489–9497, 2019.

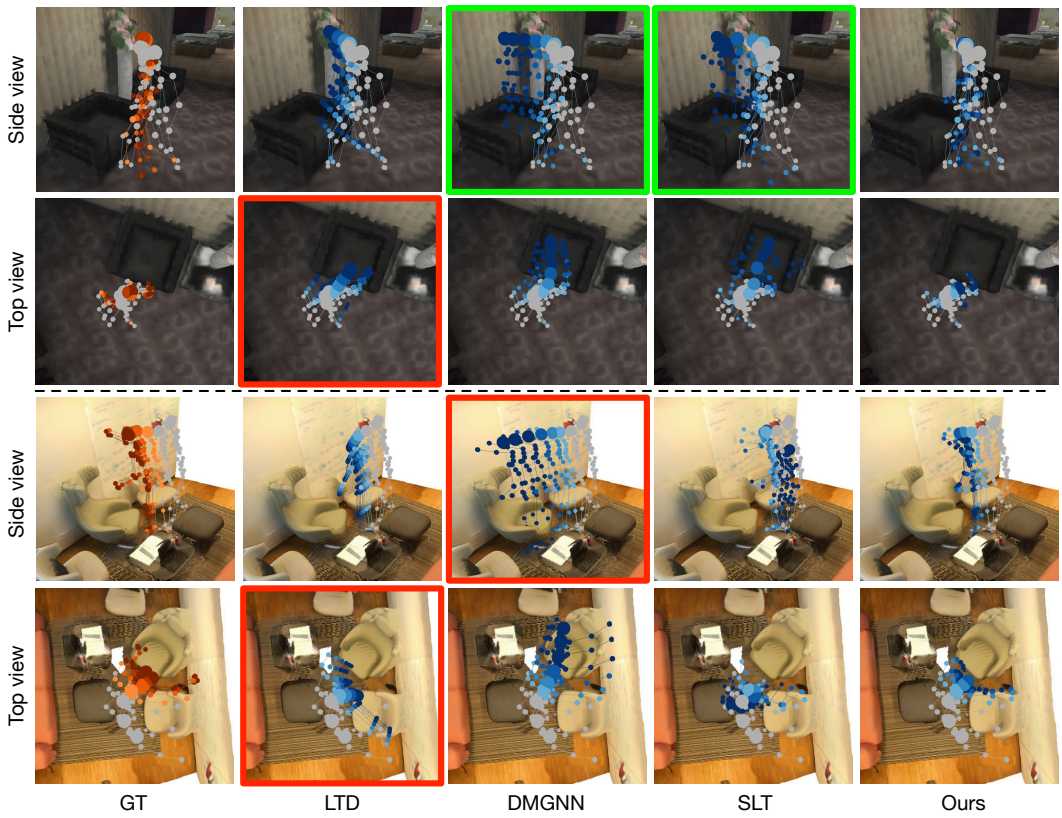


Figure 2: **Qualitative comparison to all baselines.** We highlighted motions that penetrate the scene in green and ghost motions in red. More results are in the supplementary video.

[6] 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 *CVPR*, pages 10975–10985, 2019.

[7] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, pages 9401–9411, 2021.

[8] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019.