HULC: 3D HUman Motion Capture with Pose Manifold SampLing and Dense Contact Guidance – Supplementary Material –

Soshi Shimada[1] Vladislav Golyanik[1] Zhi Li[1] Patrick Pérez[2]
Weipeng Xu[1] Christian Theobalt[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] Valeo.ai

This supplementary document provides further details on our framework (Sec. A), dataset and training details (Sec. B), the details of contact classification networks with discussions of the architecture and loss function design choices (Sec. C), and additional evaluation for the dense contact estimation network and ablation studies for the optimisations in our framework, as well as the details of the qualitative comparisons (Sec. D).

A Framework Details

In this section, we elaborate on the details of the frustum grid transform (Sec. A.1), the implementation details (Sec. A.2), the optimisation details (Sec. A.3) and the random sampling for the root translation and orientation in the sampling-based optimisation stage (Sec. A.4). We also explain, in Sec. A.5, the details of the random sampling in the kinematic pose space used in the ablation study in the main paper (Sec. 5.1).

A.1 Frustum Grid Transform

We conduct as follows the transformation from the scene point cloud $\mathbf{S} \in \mathbb{R}^{M\times 3}$, defined in the camera frame, into the frustum voxel grid $\mathbf{S}_{\mathrm{F}} \in \mathbb{R}^{32\times 32\times 256}$ whose third dimension corresponds to the discretised depth of the 3D space. Given a vertex position $p = (x, y, z)$, *i.e.*, a row of $\mathbf{S}$, in a perspective frustum space, its normalised vertex $\hat{p}$ into the cuboid space reads:

$$\hat{p} = \left(f_x \frac{x}{z}, f_y \frac{y}{z}, z \right), \tag{1}$$

where $f = (f_x, f_y)$ is the camera's focal length. The components of all points $\hat{p}$ are then suitably normalised and binned so as to build the binary occupancy grid $\mathbf{S}_{\mathrm{F}}$.

A.2 Implementations

The neural networks are implemented with PyTorch [8] and Python 3.7. We conducted the evaluations on a computer with one AMD EPYC 7502P 32 Core

Processor and one NVIDIA QUADRO RTX 8000 graphics card. The training of the contact classification networks continued until the loss convergence using Adam optimiser [5] with a learning rate 3.0×10^{-4}. Our framework runs with 25 seconds per frame excepting the computation time of SMPLify-X [9] which we use for the initial root-relative pose estimation.

A.3 Optimisation Details

For the optimization in Eq. 3 of the main paper, we use the weights $\lambda_{\text{2D}} = 1.0$, $\lambda_{\text{smooth}} = 0.01$ and $\lambda_{\text{con}} = 0.01$. For Eq. 9, λ_{sli} and λ_{data} are set to 0.05 and 0.1. In the final refinement optimisation step, we use $\lambda_{\text{data}} = 1.0$ while keeping the same weights for the other terms. Rather than using a Chamfer loss for $\mathcal{L}_{\text{con}}$ to minimise the body-environment contact vertex distance, we use the Hausdorff measure [6]; indeed, we observed that, with this measure, the reconstructed 3D motion is more robust to the false positive contact labels on the environment vertices. Note that the 2D keypoints are normalised by the image size. The joint angles are defined in radian.

A.4 Random Sampling for Root Translation and Orientation

In the pose manifold sampling-based optimisation stage, we generate candidate pose samples in the learned manifold space as described in the main paper (see Fig. 2-(b) for its schematic visualisation). For the root translation and orientation, we generate random samples around the initial translation $\boldsymbol{\tau}_{\text{opt}}$ and $\boldsymbol{\phi}_{\text{opt}}$ since they have only 3 DoF for each. Specifically, we generate samples by adding the randomly generated offsets $\Delta\boldsymbol{\tau} = \psi\boldsymbol{\varphi}_{\boldsymbol{\tau}}$ and $\Delta\text{opt} = \psi\boldsymbol{\varphi}_{\boldsymbol{\phi}}$ to $\boldsymbol{\tau}_{\text{opt}}$ and $\boldsymbol{\phi}_{\text{opt}}$, respectively; ψ is initialised to 1.0, and incremented by 1 when the solution is not found due to the hard collision constraint; $\boldsymbol{\varphi}_{\boldsymbol{\tau}} \in [-0.03, 0.03]^3$ and $\boldsymbol{\varphi}_{\boldsymbol{\phi}} \in [-0.01, 0.01]^3$ are the values generated uniformly at random. The range of $\boldsymbol{\varphi}_{\boldsymbol{\phi}}$ is kept small since even a small change of the root orientation largely modifies the 3D joint positions.

A.5 Random Sampling in the kinematic pose space for the ablation

Here, we elaborate on the details of the random sampling strategy used for the ablative study: "*our manifold sampling strategy vs. naïve random sampling with a uniform distribution in a kinematic skeleton frame*" in Sec. 5.1 in the main paper. Specifically, for the naïve random sampling, we use the random sampling for the pose parameter $\boldsymbol{\theta}_{\text{opt}} \in \mathbb{R}^{3K}$ similar to the method explained in Sec. A.4: the randomly generated offsets $\Delta\boldsymbol{\theta} = \psi\boldsymbol{\varphi}_{\boldsymbol{\theta}}$ are added to $\boldsymbol{\theta}_{\text{opt}}$ to generate the pose samples; $\boldsymbol{\varphi}_{\boldsymbol{\theta}} \in [-0.26, 0.26]^{3K}$ are the values that are uniformly generated at random.

B Dataset and Training Details

As elaborated in Sec. 4 in the main paper, we use GTA-IM [1] and PROX dataset [2] for our network training. We first pretrain our networks on the whole GTA-IM dataset using the image sequences and our body contact annotations. Lastly, we train our networks on PROX dataset with the environment contact labels obtained by us (see Sec. 4). The script to obtain the contact labels from PROX dataset and the annotated contact labels on GTA-IM dataset will be released for the future comparisons. For the evaluations, we extract test sequences from PROX and GPA [12,13] datasets. The test sequence IDs are listed in the file `test_seq_IDs.txt` in our supplement. To report the 3D per-vertex errors on GPA dataset in our main paper, we fit the SMPL-X human mesh model onto the ground-truth 3D joint keypoints. The script for this operation will also be released. In addition to the recordings in indoor scenes, PROX dataset also provides the studio recordings of the accurate 3D human shape and pose for quantitative comparison purposes. However, these recordings provide the non-contiguous sequences, hence not suitable for the evaluations of our method that requires the contiguous image sequences as one of the inputs. Therefore, we mainly report the quantitative results on GPA dataset [12,13], and qualitative results on PROX dataset. During the training, the ground-truth scene contact vertex information is once converted into the frustum voxel grid representations as elaborated on Sec. A.1. We further apply 3D Gaussian filtering to obtain the smoothed contact label signal, which helps to stabilise the network training.

C Network Details

We elaborate here on the network architectures in the dense contact estimation stage. Networks N_1 and N_3 consist of 2D-convolution-based encoder and decoder architectures. We employ Resnet-18 [4] for the encoder of N_1 without the last two layers, i.e., a fully-connected layer and an average-pooling layer. We employ U-Net [10]-based architecture for N_3 with 2 sets of down-convolution and up-convolutions blocks. Network Ω_{en} consists of 3 fully-connected layers with LeakyReLU [7] activation function. At the output layer, we use a sigmoid function instead of LeakyReLU. For the details of N_2, Ω_{bo} and the decoder of N_1, please see Figure 1.

Why This Architecture Design? Here, we discuss the architecture design choice for the environment contact estimation networks. Instead of the pixel-aligned network Ω_{en}, a 3D-convolution-based network can also be applied to obtain the voxel grid that contains per-voxel contact labels of the 3D scene. However, we observed that the 3D-convolution-based classifier network suffers from the underfitting issue during the training due to the very small number of ground-truth positive contact labels over the total number of voxels in the grid. With the pixel-aligned implicit field, we can adjust these *unbalanced* positive and negative contact labels by manipulating the sampling points in the 3D scene which we can freely control. Also, unlike the original work [11] that provides the

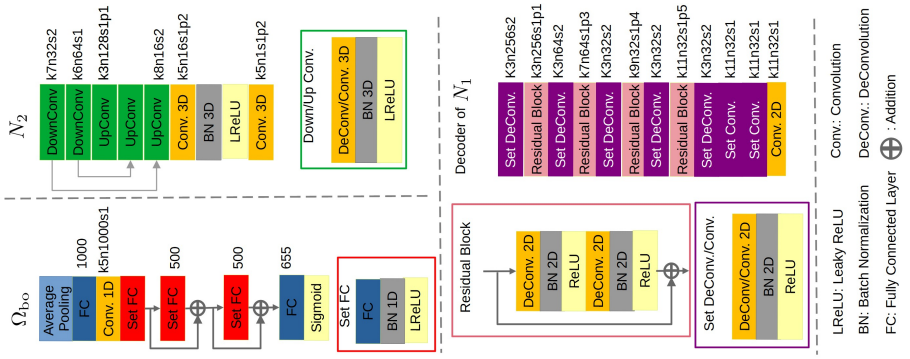

Fig. 1: The detailed network architectures for N_2, Ω_{bo} and the decoder of N_1. The numbers next to the fully-connected layers represent the output dimensionality. The numbers next to the convolution layers represent kernel size ('k'), number of kernels ('n'), size of sliding ('s') and padding size ('p'). Note that when the padding size is not shown, no padding is applied at the convolution layer.

scalar value as a depth query, we provide a one-hot vector as a depth query to Ω_{en}: we observed that it significantly reduces the loss value during the training compared to providing the scalar depth queries.

Loss Function Design (Eq. 2 in the main paper). Binary GT environment contact label ('1': contact, '0': no contact) is a very sparse signal, i.e., only a small number of voxels (˜0.01%) contain '1'. This reduces the network training stability. We observe that smoothing the environment contact labels mitigates the unbalance and enhances the training stability. Hence, with smoothing, L2-loss (not BCE) for the environment contact estimation is used. Contact labels for body are more balanced compared to the environment contacts. Therefore, we do not smooth them and use BCE loss.

D Further Evaluations and Ablations

In addition to the ablation studies and evaluations reported in our main paper, we further assess the performance of the dense contact estimation network (Sec. D.1) and the sliding loss term $\mathcal{L}_{\text{sli}}$ (Eq. 10) introduced in the main paper (Sec. D.2). Lastly, we explain the setup of the qualitative results in our video (Sec. D.3).

D.1 Contact Classifications

As HULC is the first method estimating contact labels on dense body and environment surfaces from monocular RGB and point cloud input, there are no

Table D1: Ablation study for the sliding loss term $\mathcal{L}_{\text{sli}}$.

	MPJPE [mm] ↓	sliding error [mm] ↓
Ours	**217.9**	**16.0**
Ours (w/o $\mathcal{L}_{\text{sli}}$)	220.2	18.5

other existing works that estimate the same outputs. Nonetheless, we report the performance on the GPA dataset for completeness and future reference. The precision, recall and accuracy of the body surface contact estimation are 0.22, 0.41 and 0.91, respectively. For the environment surface contact estimation, 0.045, 0.18 and 0.96, respectively. Note that these classification tasks are highly challenging, especially since the environment point cloud contains several thousands of vertices to be classified. Furthermore, GPA dataset sequences are not included in the training dataset for the contact estimation networks (see Sec.B for the training/test splits). Although it is conceivable that the reported numbers can be further improved, our framework largely benefits from the estimated contact labels and significantly reduces the 3D localisation errors as reported in Tables 2 and 3 in the main paper.

D.2 More Ablations for the optimisations

In the main paper, we performed substantial ablation studies; **(i)** 3D errors and physical plausibility measurement with the variants of our method ("Ours (w/o S)", "Ours (w/o R)" and "Ours (w/o SR)") in Tables 2 and 4, **(ii)** with and w/o (denoted as "Baseline") contact loss term $\mathcal{L}_{\text{con}}$ (Eq. 6) in Table 3, **(iii)** experiments with different number of samples in the sampling-based optimisation stage (Fig. 3-a), **(iv)** experiments with different number of iterations in the sampling-based optimisation (Fig. 3-b), and **(v)** with and w/o the confidence merging strategy (Eq. 7) in Sec. 5.1.

For the completion, we report the ablative study for the sliding loss term $\mathcal{L}_{\text{sli}}$ (Eq. 9) used in our optimisations. In Table D1, we report MPJPE and sliding error $e_{\text{sli.}}$ measured in a world frame for our full framework ("Ours") and our framework w/o the sliding loss term ("Ours (w/o $\mathcal{L}_{\text{sli}}$)"). The sliding error $e_{\text{sli.}}$ is measured by computing the average of the drift of the contact vertex on the human surface, based on the assumption that contact positions in the scene are not moving (*i.e.,* zero velocity). This is a reasonable assumption since most of the contact positions in daily life in a static scene are static contacts, which is also the case with our evaluation dataset; GPA dataset [12,13].

With the sliding loss term, our framework reduces the sliding error by ∼14% compared to w/o $\mathcal{L}_{\text{sli}}$. Notably, integration of $\mathcal{L}_{\text{sli}}$ reduces the 3D joint error (MPJPE) by 1% as well. Note that the ablation studies of the loss terms (*e.g.,* 2D reprojection and temporal smoothness terms) other than $\mathcal{L}_{\text{sli}}$ and $\mathcal{L}_{\text{con}}$ (Tab. 3 in the main paper) are not of interest as those are already widely used in many wroks in video-based monocular 3D human MoCap and their significance is already well known.

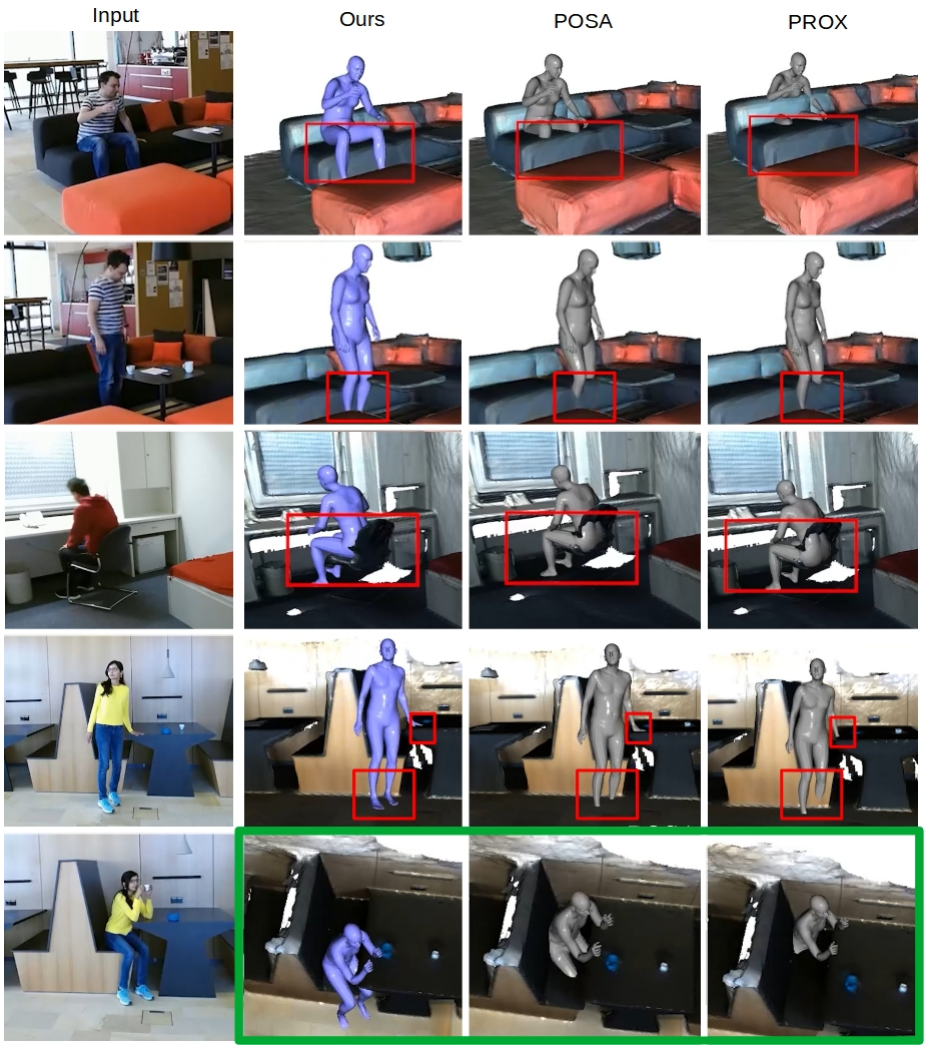

Fig. 2: Qualitative comparison of our HULC *vs* existing scene-aware RGB-based methods on PROX [2] dataset. Our method shows significantly mitigated collisions thanks to our novel sampling-based optimisation, which handles the severe body-environment penetrations in a hard manner (red rectangles). We also show the results from a top view (green rectangle). Thanks to the contact-based optimisation using the estimated dense contacts on the body surfaces and the environment, our estimated 3D global root positions are significantly more accurate compared to the previous methods.

D.3 Qualitative Comparisons

In our video, we compare our HULC with the SOTA methods PROX[2] and POSA[3] from the RGB-based algorithm class. From the RGB-D based algorithm class, we choose PROX-D [2] and LEMO[14]. For a fair comparison, Gaussian smoothing is applied to all those related methods. In Fig. 2, we show comparisons of our method with other RGB-based scene-aware methods POSA[3] and PROX [2]. Our method shows physically more plausible interactions with the environment than the others. We also visualise the result from a bird's eye view to show the significance of the contact-based optimisation, which contributes to substantially more accurate global translation estimation than other related methods (green rectangle).

References

1. Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision (ECCV) (2020) 3
2. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (ICCV) (2019) 3, 6, 7
3. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: Computer Vision and Pattern Recognition (CVPR) (2021) 7
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR) (2016) 3
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations, ICLR (2015) 2
6. Knauer, C., Löffler, M., Scherfenberg, M., Wolle, T.: The directed hausdorff distance between imprecise point sets. In: International Symposium on Algorithms and Computation (ISAAC) (2009) 2
7. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning (ICML) (2013) 3
8. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS) (2019) 1
9. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2019) 2
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015) 3
11. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: International Conference on Computer Vision (ICCV) (2019) 3
12. Wang, Z., Chen, L., Rathore, S., Shin, D., Fowlkes, C.: Geometric pose affordance: 3d human pose with scene constraints. In: Arxiv (2019) 3, 5
13. Wang, Z., Shin, D., Fowlkes, C.: Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In: European Conference on Computer Vision Workshop (ECCVW) (2020) 3, 5
14. Zhang, S., Zhang, Y., Bogo, F., Marc, P., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: International Conference on Computer Vision (ICCV) (Oct 2021) 7