HMD-NeMo: Online 3D Avatar Motion Generation From Sparse Observations

Sadegh Aliakbarian Fatemeh Saleh David Collier Pashmina Cameron Darren Cosker

Microsoft Mixed Reality & AI Lab, Cambridge, UK

Qualitative Results. We highly encourage readers to see the supplementary video submitted with this paper, containing multiple examples that show the output of HMD-NeMo given the HMD signal in hand tracking scenario.

1. Implementation Details

To train HMD-NeMo, we utilize the Adam optimizer [3] with a batch size of 256 and a learning rate of $1e^{-3}$. We follow [2] and train HMD-NeMo with sequences of length 40 frames, however, our approach can be used to generate sequences of arbitrary length at inference time. To optimize HMD-NeMo's prediction, we use limited-memory BFGS optimizer [4], with a history size of 10, learning rate of 1, and Strong-Wolfe line search function [5]. We only optimize the upper body pose parameters as well as the global root trajectory as the observations (head and hands) represent upper body only. In the rest of this section, we describe the detailed design of each component of HMD-NeMo.

Head and hand embedding module. This module comprises four shallow MLPs per each 6-DoF of the input: a MLP to compute the rotation representation, a MLP to compute the translation representation, a MLP to compute the rotational velocity representation, and finally a MLP to compute the positional velocity representation. Each MLP is a single `Linear` layer followed by `LeakyReLU` non-linearity. Each MLP maps its input (either rotation in 6D or translation in 3D) to a vector of size 32 in the latent space. For each 6-DoF representation (i.e., head, left hand, right hand, left hand in the head space, and right hand in the head space), the result of the four MLPs are then concatenated together to form a vector of size 128.

SpatioTemporal encoder (STAE). This module comprises two sub-modules: a GRU-based module to encode temporal information and a transformer-based module to encode spatial information. Given the embedding representation of each input 6-DoF (of size 128), we consider a single-layer `GRU` with the hidden size of 256 to process each input signal temporally. The hidden state of each GRU cell is updated given its input and the previous hidden state. For each GRU, we initialize the hidden state at time $t = 0$ with a MLP (a `Linear` layer followed by `Tanh` non-linearity) that gets as input the head embedding at time $t = 0$ (head is considered the reference joint and it is always visible) and computes the initial hidden state. For each input representation, we have a separate GRU layer and a separate hidden state initialization MLP. Since we have five 6-DoFs in the input signal, thus we compute and update five separate hidden states of the GRU. Such hidden states encompass the temporal information about each component of the input separately. At each time-step, these hidden states are then used as the input to a transformer encoder to learn how these temporal features are spatially correlated to each other. Specifically, we use 4 layers of transformer encoder, each with 4 attention heads and a feed-forward hidden dimension of 512.

Temporally adaptable mask tokens (TAMT). In order to take care of missing observations for hands, where computing the hand embedding representations is not feasible, we introduce TAMT, as described in the main paper. For each hand, TAMT contains a base MLP (two layers of `Linear-LeakyReLU`) which gets as input the concatenation of the head representation and the corresponding hand representation, computed by the transformer encoder at time t. The output of the base MLP is then passed as input to two separate MLPs: a MLP (a single `Linear` layer followed by `LeakyReLU` non-linearity), called ToToken, that computes a vector of size 128 (same size as the output of head and hand embedding module) that produces TAMT features for time $t + 1$, and a MLP (a `Linear` layer, followed by `LeakyReLU` non-linearity, followed by another `Linear` layer), called Forecaster, that computes/forecasts the 6-DoF of the corresponding hand in time $t + 1$. The base MLP produces a feature vector of size 256 and Forecaster module's intermediate hidden dimension is also 256. For the very initial time-step, if a hand observation is missing, we use a learned parameters for TAMT (learned via `nn.Parameters`).

2. Additional Ablation Studies

2.1. Robust energy term

Why we need a robust energy term? As described in the main paper, once trained, HMD-NeMo is capable of generating high fidelity and plausible human motion given only the HMD signal. However, as is typical of learning-based approaches, the direct prediction of the neural network does not precisely match the observations i.e., the head and hands, even if it is perceptually quite close. To close this gap between the prediction and the observation, optimization can be used. This adjusts the pose parameters to minimize an energy function of the form $\mathcal{E} = \mathcal{E}_{data} + \mathcal{E}_{reg}$, where $\mathcal{E}_{data}$ is the energy term that minimizes the distance between the predicted head and hands to the observed ones, and $\mathcal{E}_{reg}$ is additional regularization term(s). To define the data energy term, we define the residual $\mathcal{R} = \sum_{j \in \{h,l,r\}} (x_j - \hat{x_j})$, i.e., the difference between the predicted head/hand joint to that of the observation. Given $\mathcal{R}$, a typical, non-robust data energy term could be written as $\mathcal{E}_{nr} = \mathcal{R}^2$, i.e., the L2 loss. This suits the MC scenario perfectly, where head, left hand, and right hand are *always* available. But this energy term may be misleading in HT scenario where hands are going into and out of FoV often, and thus leading to abrupt pose changes when hands appear back in the FoV. Specifically, consider that the right hand was out of FoV for a relatively long period of time up to time t and the model has predicted what the right arm motion could be like for this period. Then, at time t, the right hand comes back to the FoV and thus we have an observed right hand signal. While the motion generated by the model is plausible, the predicted right hand may end up in a completely different location from the newly observed right hand. If we use the data energy term $\mathcal{E}_{nr}$ to minimize the total energy during optimization, we end up in an abrupt jump in the right arm pose from time $t-1$ to time t. While this guarantees high fidelity, i.e., hands in the correct position once observed, it adversely affects the perceptual experience of generating temporally smooth and coherent motion.

Parameters of robust energy term. Inspired by the general form a robust loss function [1], in this paper we use such technique to define our robust energy term for the optimization in the hand tracking scenario. In hand tracking scenario, where hands may appear outside of the field of view of the HMD, a non-robust energy term, e.g., L2, is ideal for the fidelity (hand poses appearing accurately when there is a hand observation), while may not be ideal for plausibility, as a result of abrupt jumps when a new hand signal is observed. This is why a robust alternative

$$\mathcal{E}_r(\mathcal{R}, a, b, c) = b\frac{|a-2|}{a}\left(\left(\frac{(\frac{\mathcal{R}}{c})^2}{|a-2|}+1\right)^{(\frac{a}{2})}-1\right) \quad (1)$$

is used when plausibility in the generated motions is a priority. As described in the main paper, the values of the hyper-parameters a, b, and c affect the shape and thus the behaviour of the energy term. Particularly, such hyper-parameters determine (1) what range of values of $\mathcal{R}$ should be considered outlier, and thus not being penalized strongly, and (2) what is the penalty strength for the inliers and outliers. The effect of each parameter is visualized in Fig. 1. Particularly, parameter a determines the strength of the

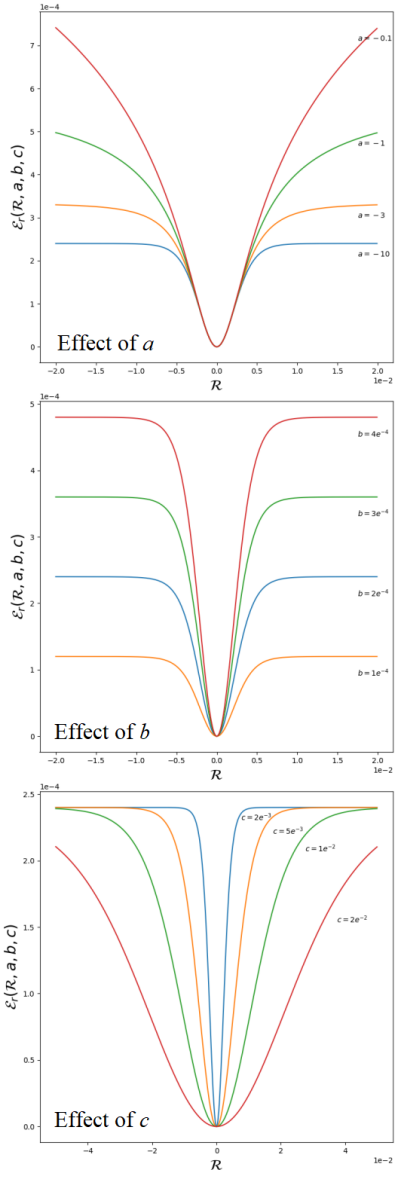

Figure 1. Effect of each hyper-parameter in determining the shape of the robust energy term, Eq. 1.

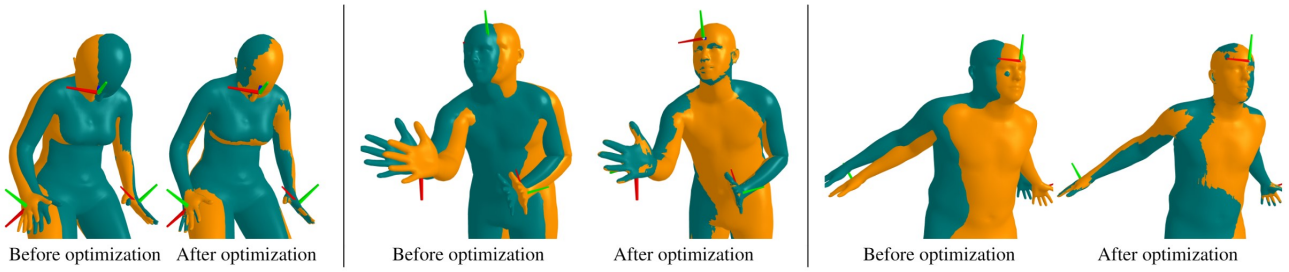

Figure 2. Illustration of the effect of optimizing HMD-NeMo predictions (predictions are shown in teal and GT in orange, overlaid). The initial model prediction is relatively accurate, but just a single optimization iteration improves the head and hands prediction substantially.

Setting	Metric	Motion Controllers	Hand Tracking
Full body	MPJPE ↓	2.07	2.48
	MPJVE ↓	26.07	31.30
Upper body	UB-MPJPE ↓	1.87	2.28
	UB-MPJVE ↓	24.26	29.76
Lower body	LB-MPJPE ↓	2.56	2.80
	LB-MPJVE ↓	30.77	33.03
Head & Hands	HH-MPJPE ↓	0.88	1.72
	HH-MPJVE ↓	13.83	28.74

Table 1. Per body part evaluation. Note these results represent the HMD-NeMo prediction prior to optimization.

penalty as outliers go further from inlier region. Parameter b determines the value of the loss at which it considers outlier. Parameter c determines width at which we consider $\mathcal{R}$ as inlier. In our experiments, $a = -10$, $b = 2e^{-4}$, and $c = 2e^{-3}$. Numbers on the plots of Fig. 1 are best seen when zoomed in. Fig. 2 illustrates multiple examples before and after only 1 iteration of optimization.

2.2. Evaluation on various body parts

In Table 1 we compare the MC and HT scenarios, and break down the errors for various body parts. Since the HMD signal represents the upper body, the contribution of lower body joints towards the error (both MPJPE and MPJVE) is larger than that of the upper body joints. As expected, head and hand errors are relatively low since HMD signals represent head and hands. Also, as expected, the results in Table 1 demonstrate that the motion prediction task in HT scenario is more difficult than in MC. Despite being user-friendly, HT scenario has not been well-explored by the community yet due to the technical difficulty of motion prediction in this setting.

3. Additional Qualitative Results

In this section, we first provider results of HMD-NeMo (see Fig. 3 to Fig. 6) and then provide more qualitative comparisons to the state-of-the-art approach [2] (see Fig. 7 to Fig. 28). Note that none of the results are cherry picked. Note that all results are color-coded based on the distance between the predicted vertices and that of the ground truth. Dark blue vertices represent predictions that are very close to the ground truth and yellow vertices are further away from ground truth.

In comparison to [2], both our approach and the baseline performs reasonably well in predicting the upper-body mainly due to the fact that HMD signal is a strong signal about the upper body pose. Typically, HMD-NeMo does a relatively better job at predicting more plausible lower body motion.

References

[1] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019. 2

[2] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. AvatarPoser: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of European Conference on Computer Vision*. Springer, 2022. 1, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations, ICLR*, 2015. 1

[4] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 1

[5] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006. 1

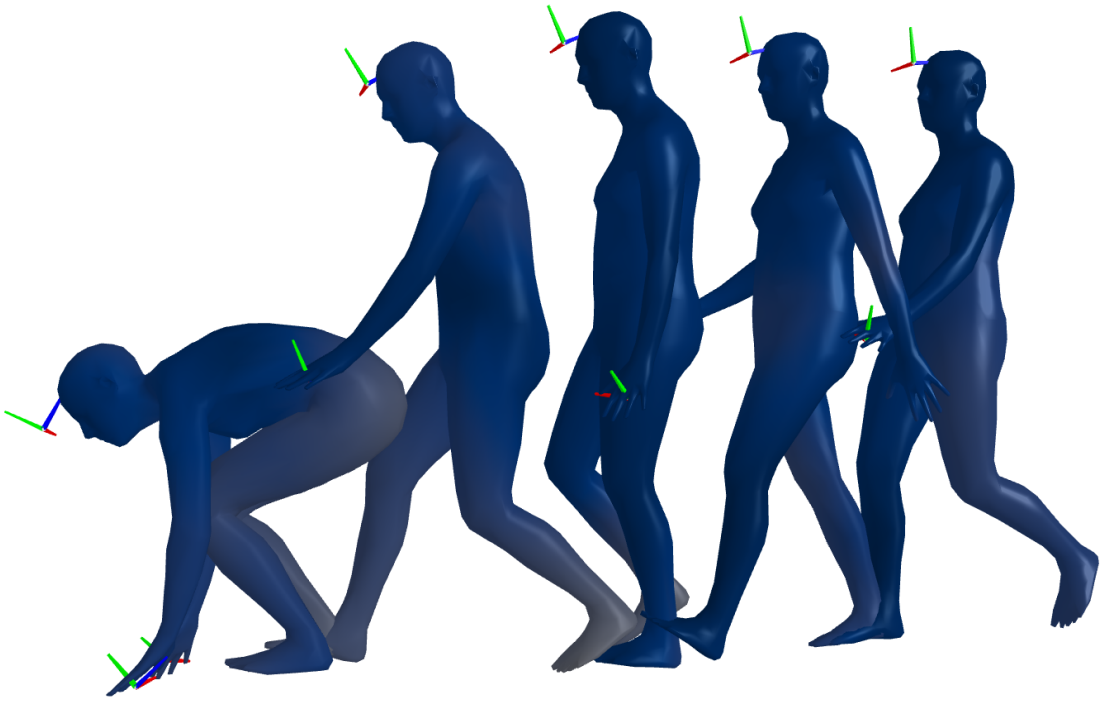

Figure 3. Qualitative results of HMD-NeMo in HT scenario.

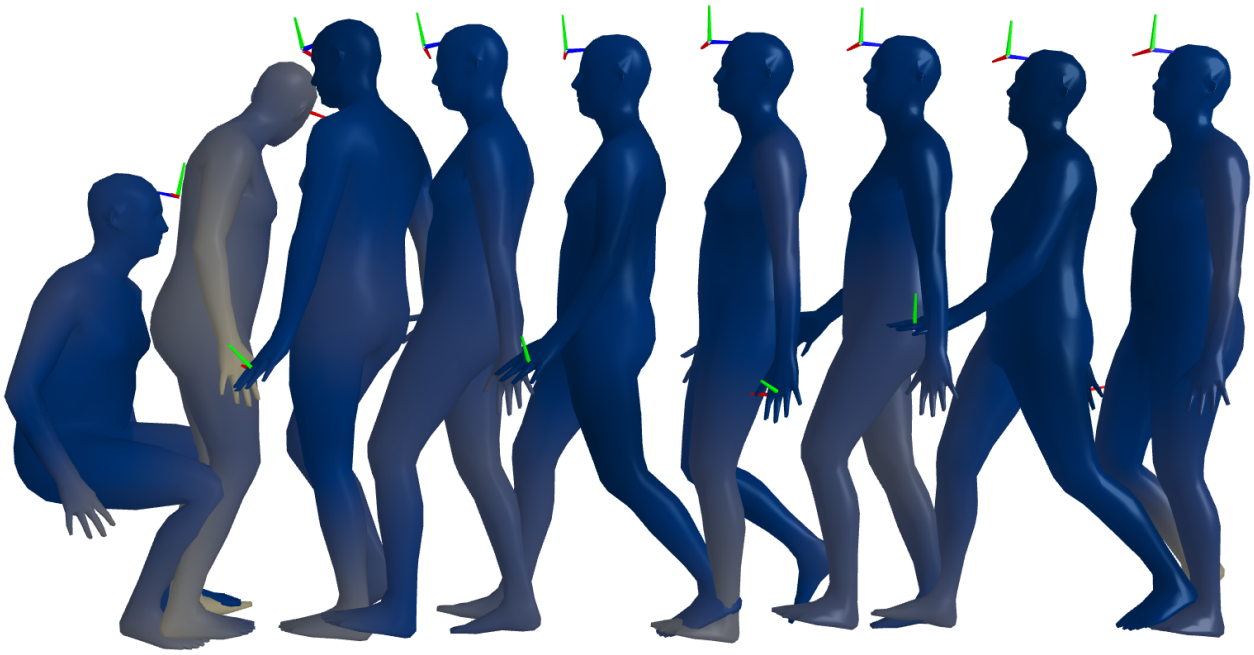

Figure 4. Qualitative results of HMD-NeMo in HT scenario.

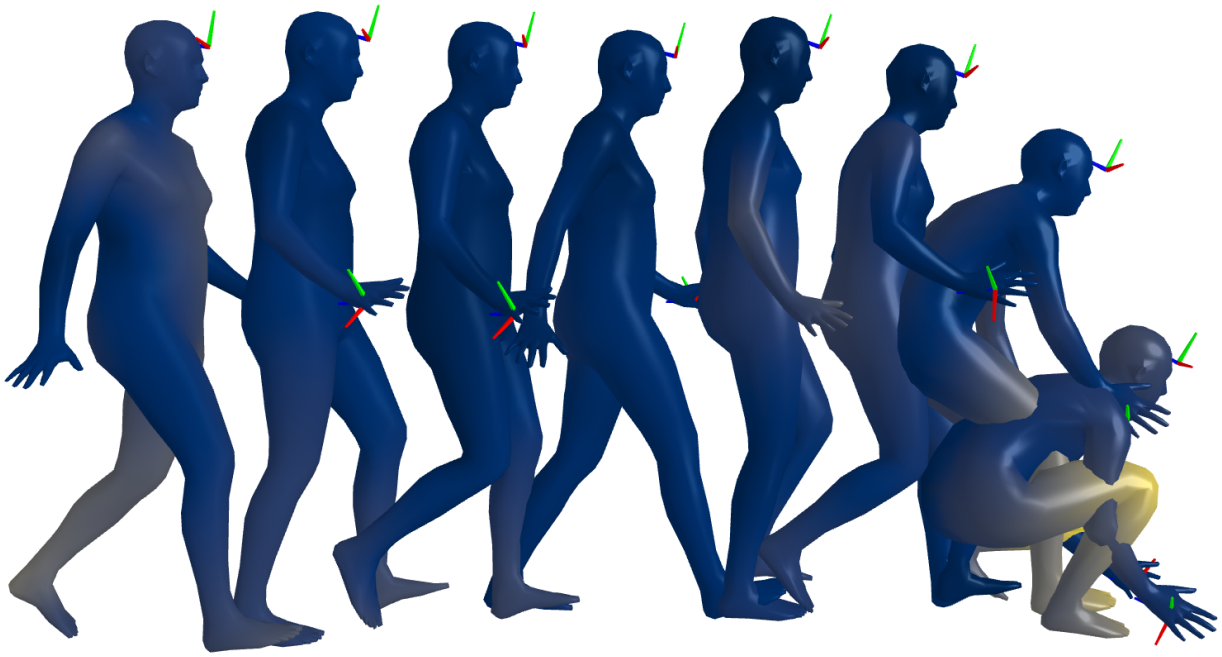

Figure 5. Qualitative results of HMD-NeMo in HT scenario.

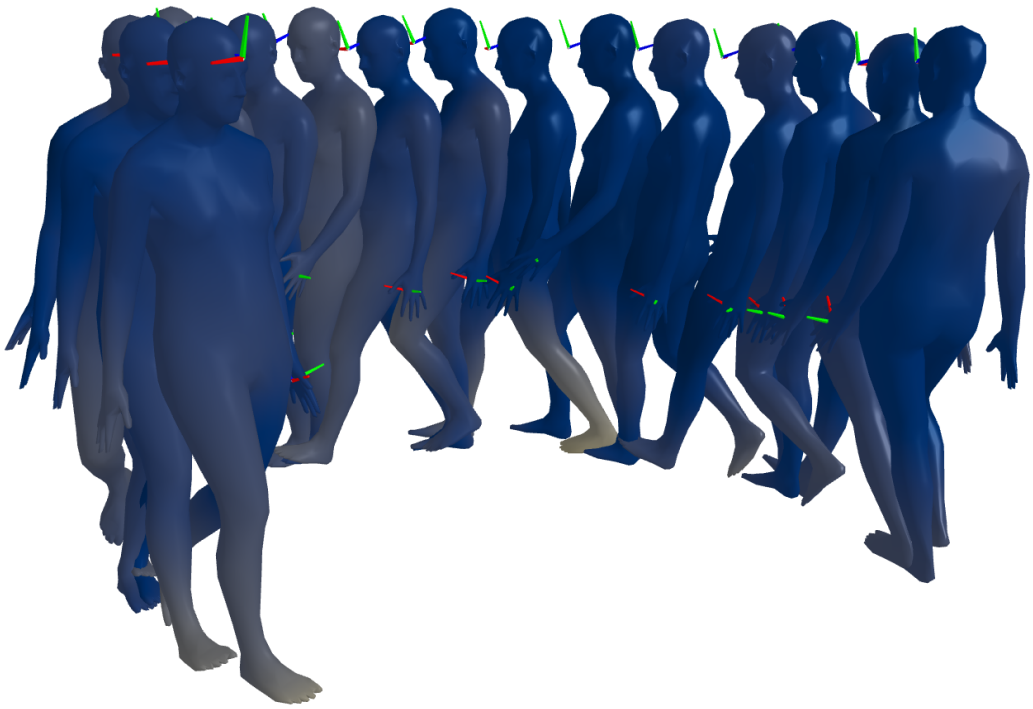

Figure 6. Qualitative results of HMD-NeMo in HT scenario.

Figure 7. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

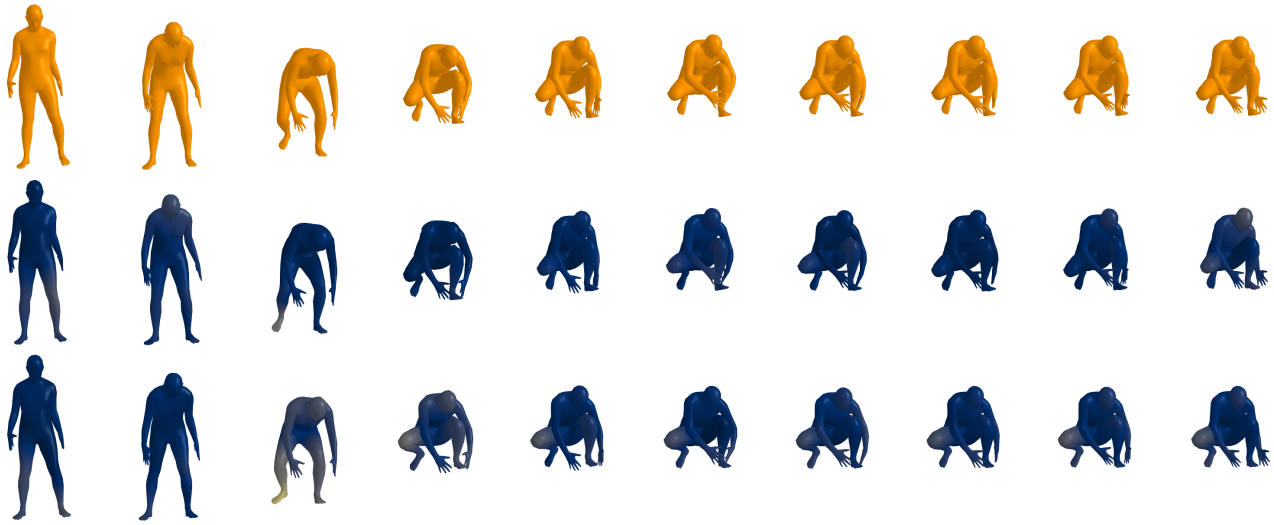

Figure 8. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 9. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 10. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

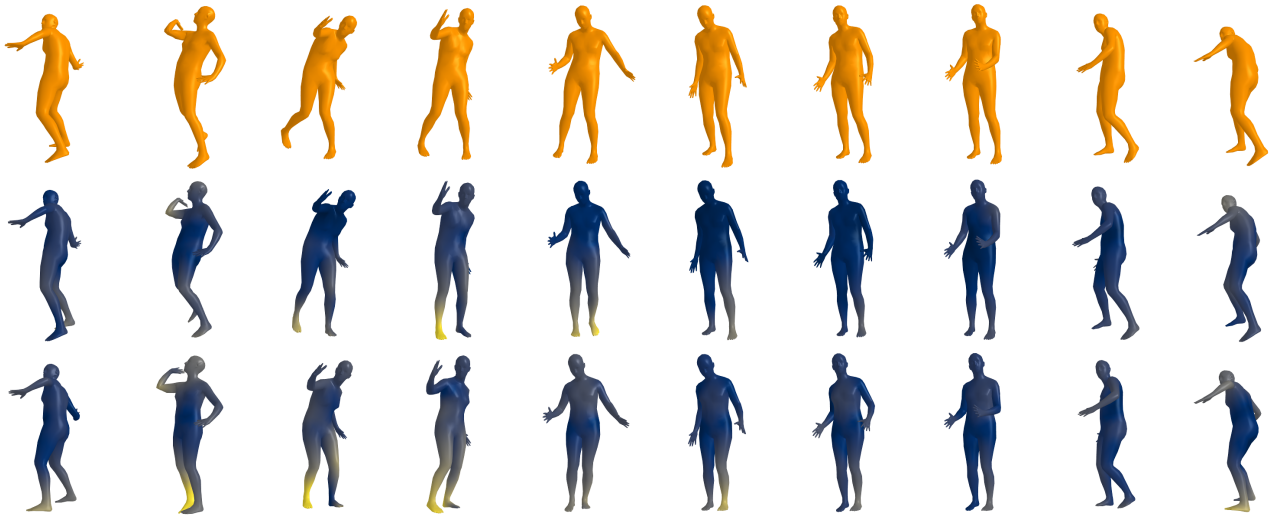

Figure 11. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

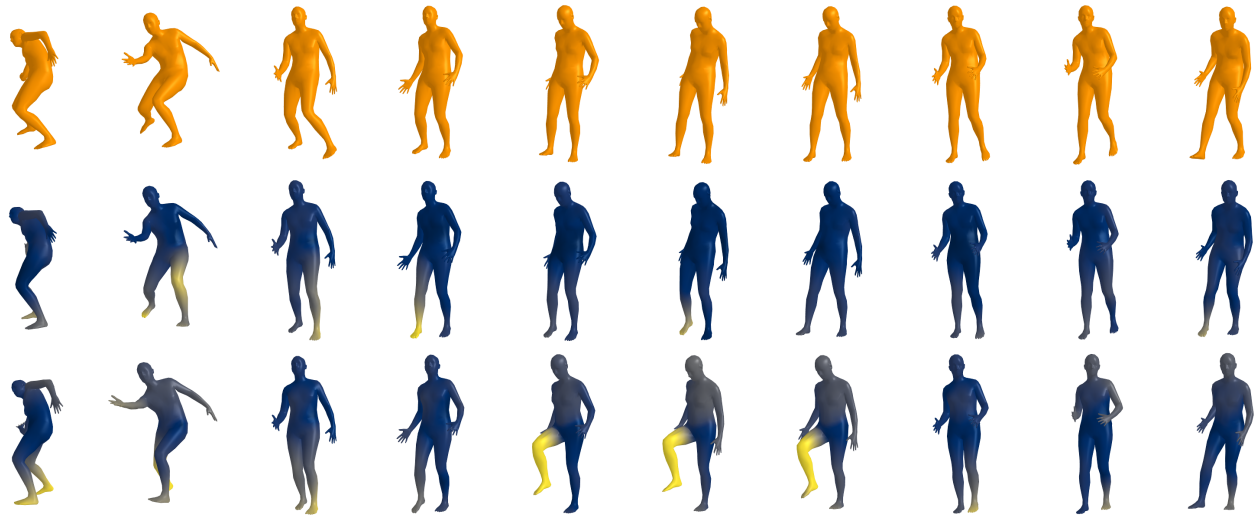

Figure 12. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

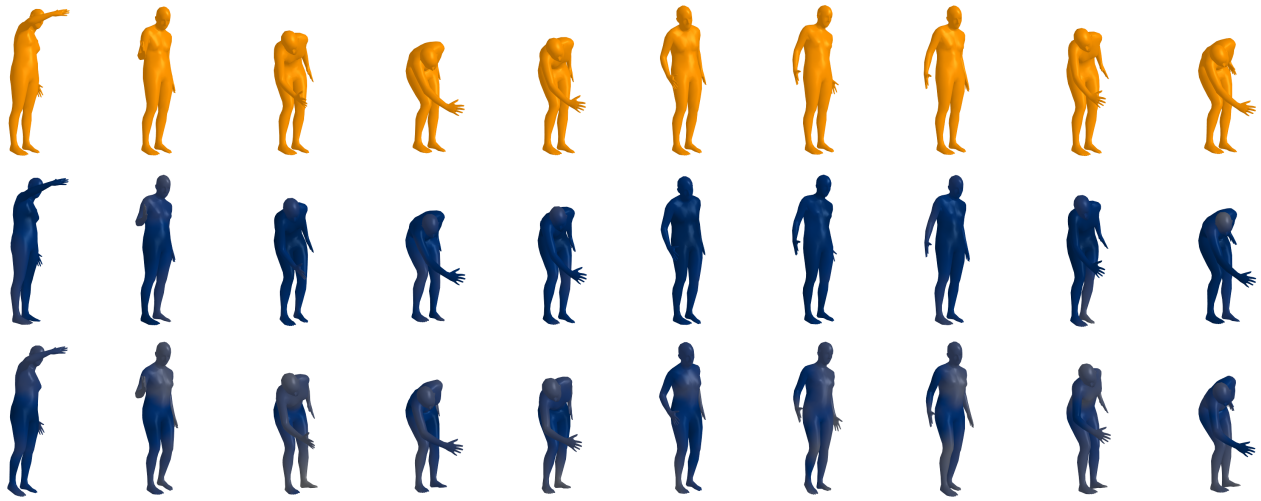

Figure 13. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

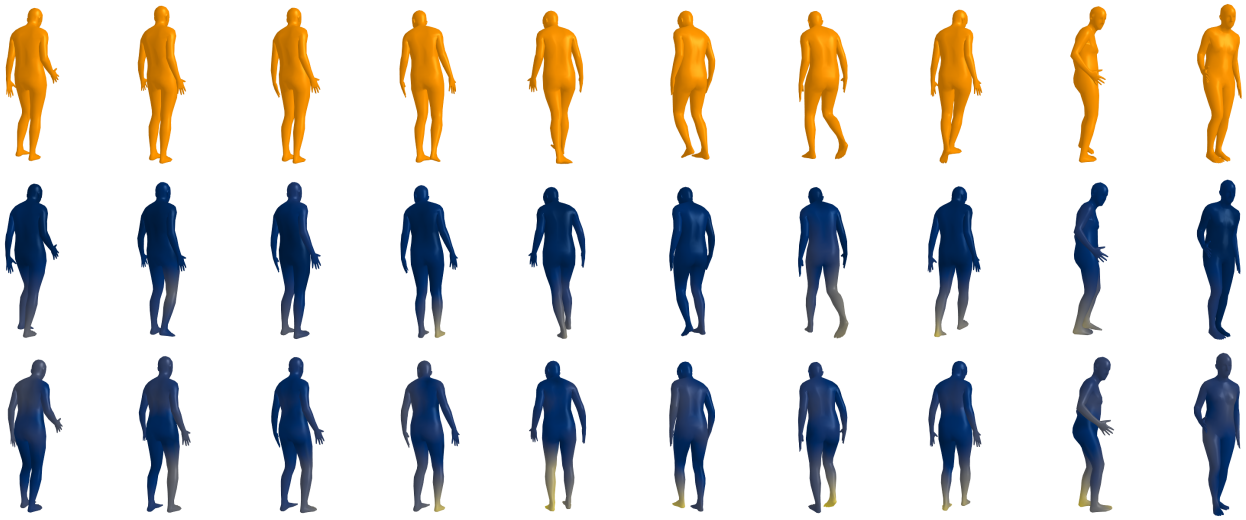

Figure 14. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 15. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

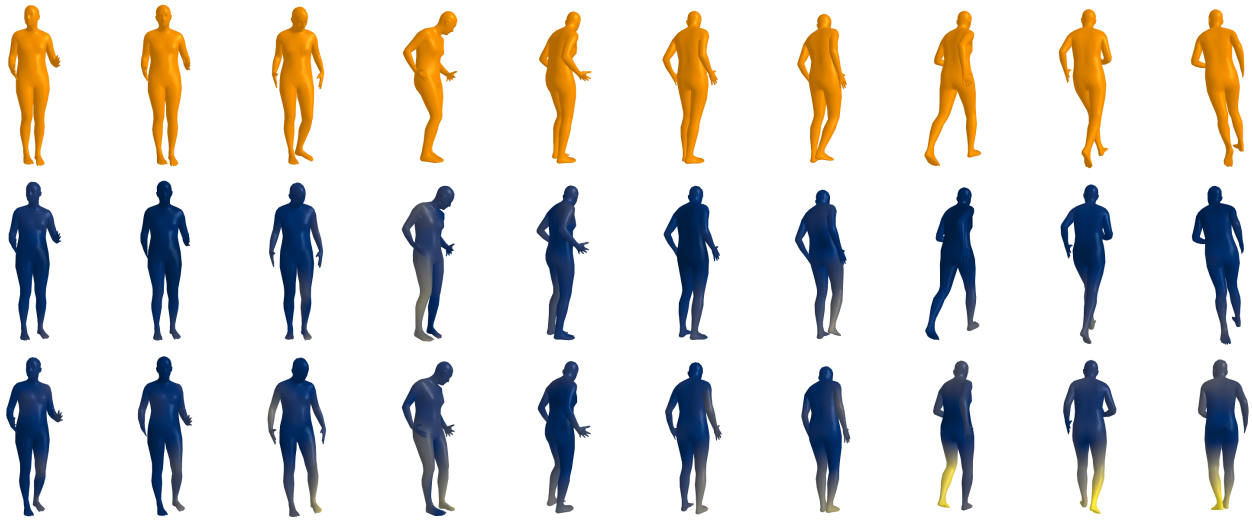

Figure 16. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

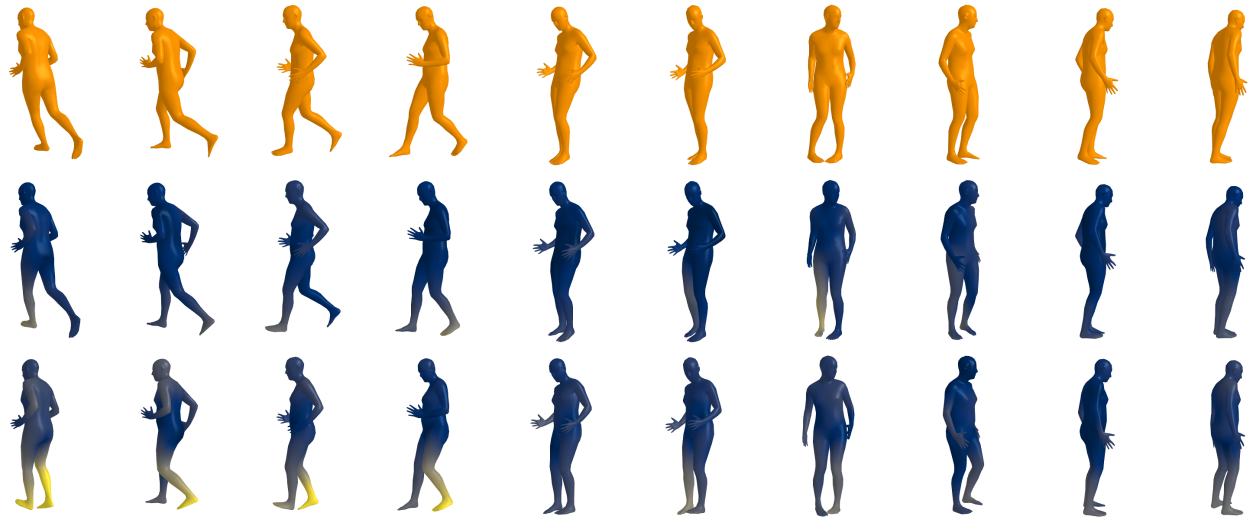

Figure 17. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 18. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

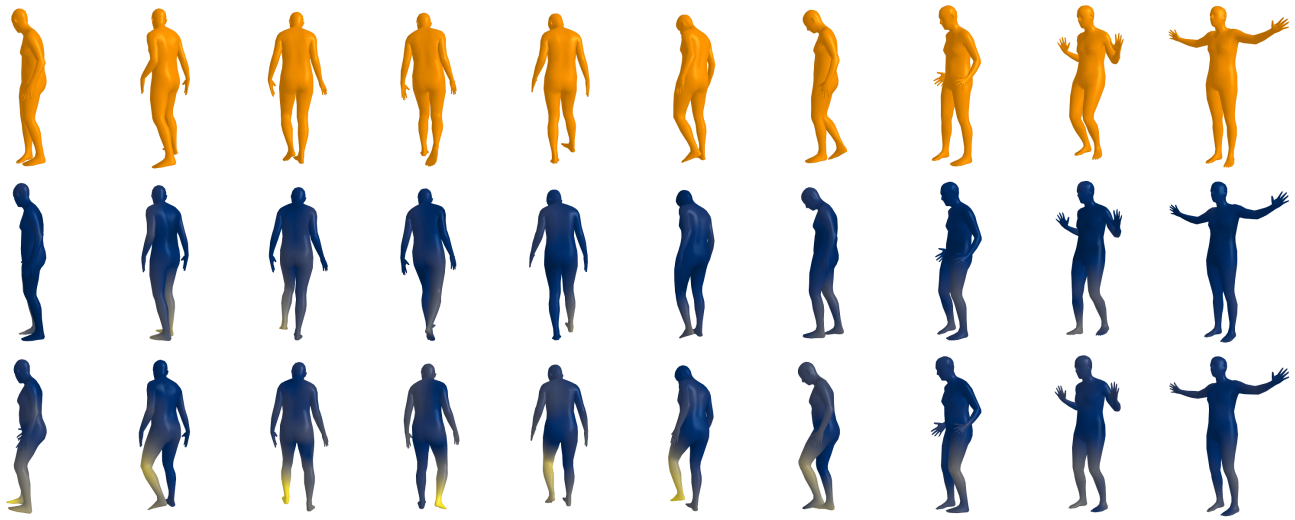

Figure 19. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

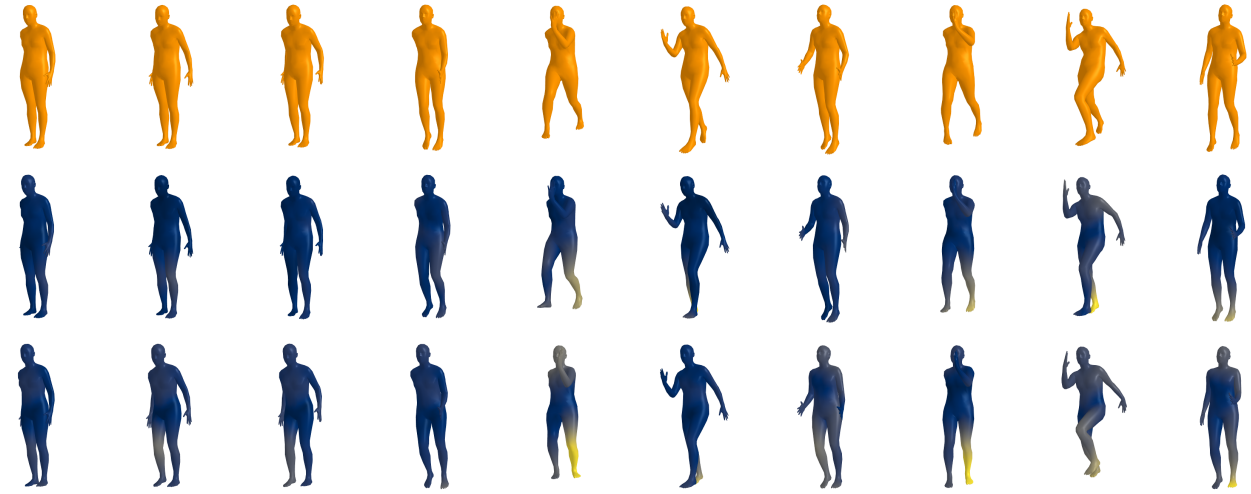

Figure 20. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

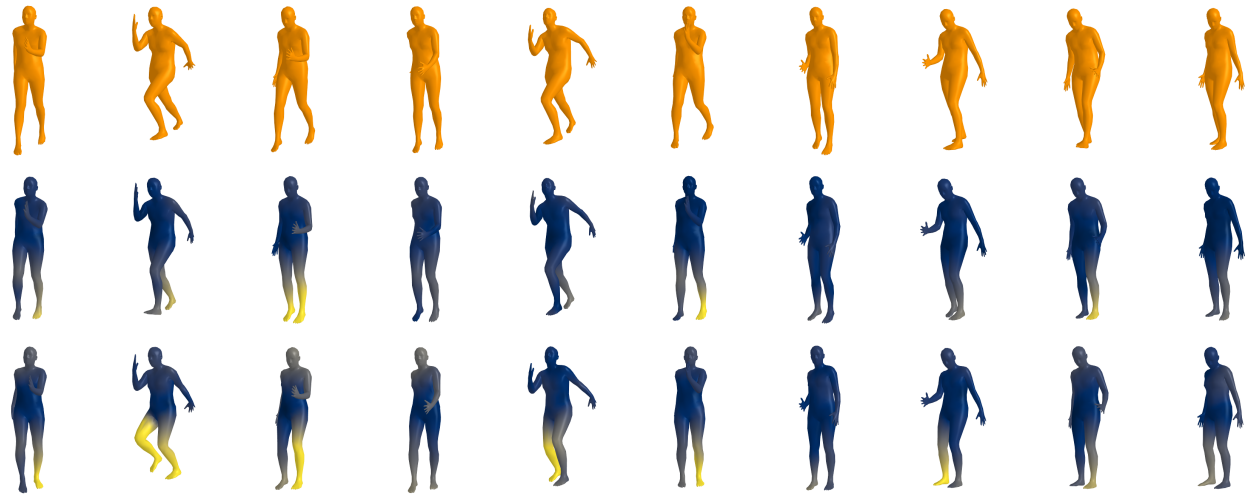

Figure 21. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 22. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

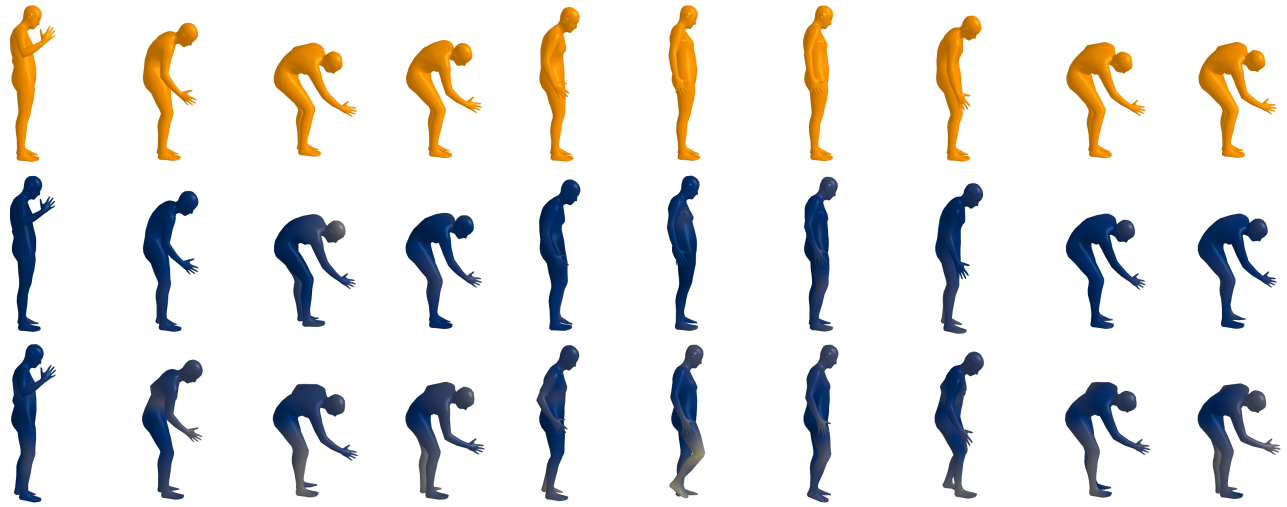

Figure 23. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

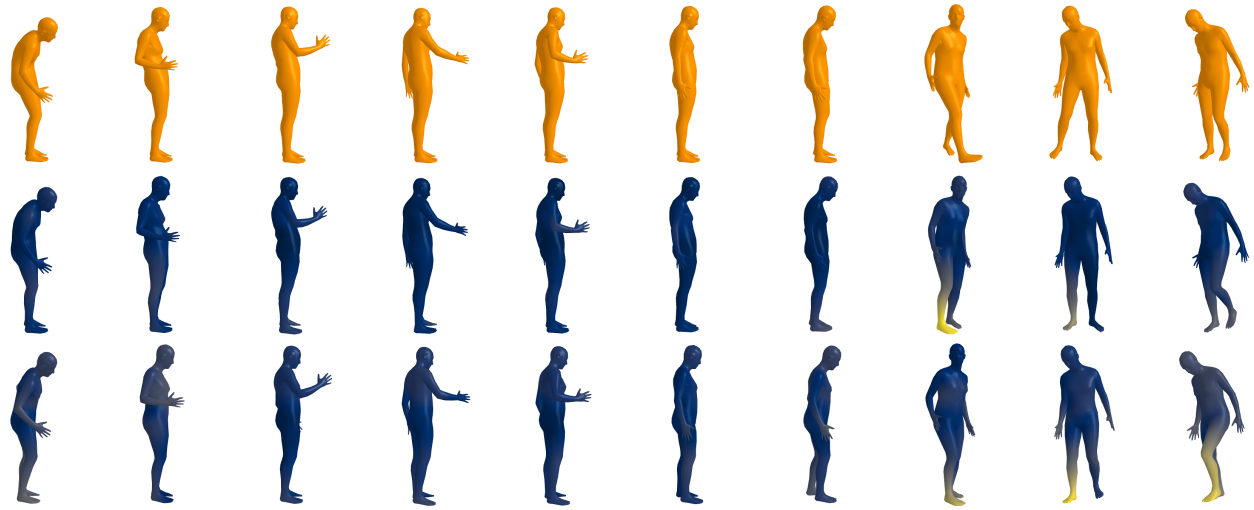

Figure 24. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 25. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 26. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

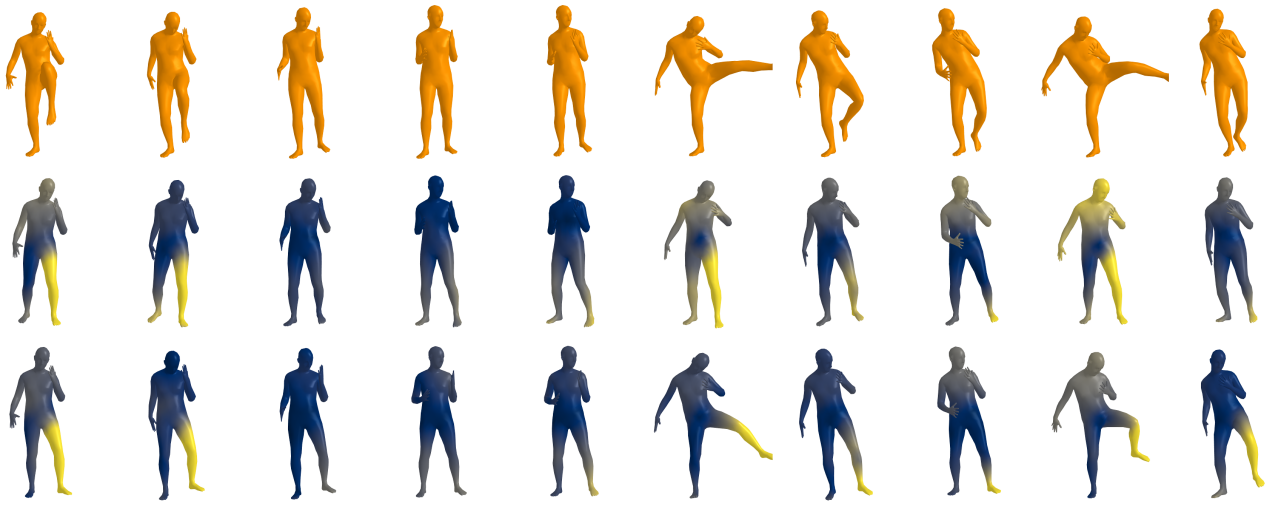

Figure 27. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].

Figure 28. Qualitative comparison to the state of the part method [2] in MC scenario. **Top**: Ground truth in orange, **Middle**: HMD-NeMo, **Bottom**: Jiang et al. [2].