Supplementary Material
PINA: Learning a Personalized Implicit Neural Avatar from a Single RGB-D Video Sequence

Zijian Dong*[1] Chen Guo*[1] Jie Song†[1] Xu Chen[1,2] Andreas Geiger[2,3] Otmar Hilliges[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen
[3]University of Tübingen

In this **supplementary document**, we provide additional materials to supplement our main submission. In Sec. 1, we provide further implementation details of our proposed method. Sec. 2 explains details regarding the exact implementation of baseline methods. Furthermore, in Sec. 3 we provide more results, including ablation studies on effects of the energy functions (Sec. 3.1), additional qualitative results (Sec. 3.2), comparisons on real data (Sec. 3.3). More comparisons with traditional fusion-based methods (Sec. 3.4) and SCANimate (Sec. 3.5) are also supplemented. Finally, we discuss our limitations and potential negative societal impacts in Sec. 4. In the **supplementary video**, we show reconstruction and animation demos of our method on real-world data.

1. Implementation Details

1.1. Architecture

We leverage MLPs to represent the shape network and skinning network. The shape network includes 8 blocks, each of which consists of one fully connected layer, a weight normalization layer [16] and a softplus activation layer [7]. Each fully connected layer contains 256 neurons. The pose condition $\mathbf{p}$ is obtained by concatenating all axis angles and is passed through a fully connected layer to reduce its dimension to 8. We concatenate this pose feature and the output feature of the fourth block as the input of the fifth block of this network. We use geometric initialization [2] for the shape network's weights. We apply positional encoding with 4 frequency components to the input points to better model high-frequency details [14]. Since the skinning weights are smooth and thus don't require a large network to predict, we only use 4 blocks containing 128 neurons without conditioning on pose $\mathbf{p}$.

1.2. Iterative Root Finding

To find the set of canonical correspondences of the deformed point $\mathbf{x}_d$, we follow [6] to define them as the root $\mathbf{x}_c$ of the following linear blending constraint:

$$\sum_{i=1}^{n_b} w_c^i \mathbf{B}_i x_c - x_d = f(\mathbf{x_c}, \mathbf{B}, \mathbf{x_d}) = f(\mathbf{x}_c) = 0 \quad (1)$$

where $\mathbf{w}_c = \{w_c^1, ..., w_c^{n_b}\} = f_w(\mathbf{x}_c)$ represents the learned skinning weights for $\mathbf{x}_c$ and $\mathbf{B}$ is the bone transformation matrix derived from the pose $\mathbf{p}$. Only $\mathbf{x}_c$ is unknown in this equation.

To find the solution to equation $f(\mathbf{x}_c) = 0$, we leverage Broyden's method [4] for our correspondence search. In our experiments, we set the maximum number of update steps to 50 and the convergence threshold to 10^{-5}. Finally, we choose the top-k candidates with the lowest errors as our canonical candidates by conducting random re-initialization. k is empirically set to 9 in our experiment.

1.3. Training Details

We train our network using the Adam optimizer [9], with a learning rate of $r_{\text{sdf}} = 10^{-4}$ for the implicit canonical shape network and $r_w = 10^{-6}$ for the canonical skinning field network. The other Adam hyper-parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A model is trained for 12h on a single NVIDIA RTX TITAN GPU.

1.4. Pose Initialization

We initialize body poses by fitting SMPL model [10] to raw RGB-D observations. We formulate this process as a per-frame optimization problem. Specifically, this is achieved by minimizing the following loss functions over the SMPL parameter $\mathbf{P}_i$ consisting of the pose $\mathbf{p}_i$, shape $\boldsymbol{\beta}_i$ and global translation $\mathbf{t}_i$:

$$\mathcal{L}^i(\mathbf{P}_i) = \mathcal{L}^i_{p2s}(\mathbf{P}_i) + \mathcal{L}^i_{DP}(\mathbf{P}_i) + \mathcal{L}^i_T(\mathbf{P}_i) + \mathcal{L}^i_{\text{prior}}(\mathbf{P}_i) \quad (2)$$

*Equal contribution
†Corresponding author

This loss formulation considers the point-to-surface loss $\mathcal{L}^i_{p2s}$, dense correspondence loss $\mathcal{L}^i_{DP}$, temporal pose stability loss $\mathcal{L}^i_T$ and pose plausibility $\mathcal{L}^i_{\text{prior}}$. We now introduce these loss terms in more detail:

Point-to-Surface $\mathcal{L}^i_{p2s}$ represents a point-to-surface loss for frame i, measuring the sum of the distances from the points $\mathbf{x}_d$ to the SMPL body surface. We uniformly downsample the point clouds to retain only n_{pc} points, and compute the energy function via:

$$\mathcal{L}^i_{p2s}(\mathbf{P}_i) = \sum_{j=1}^{n_{pc}} \frac{P2S(\mathbf{x}^j_d, f_{smpl}(\mathbf{P}_i))}{n_{pc}} \tag{3}$$

where $P2S(\cdot)$ calculates the ℓ_2 point-to-surface distance and $f_{smpl}(\cdot)$ denotes the forward function that regresses the SMPL body surface from the SMPL parameters.

Dense Correspondence $\mathcal{L}^i_{DP}$ depicts a 3D per-pixel DensePose [8] energy term where we minimize the distance between the point $\mathbf{x}_d$ and its corresponding 3D position on SMPL surface $\mathbf{x}_{smpl}$ estimated by DensePose:

$$\mathcal{L}^i_{DP}(\mathbf{P}_i) = \sum_{j=1}^{n_{DP}} \frac{P2P(\mathbf{x}^j_d, \mathbf{x}_{smpl})}{n_{DP}} \tag{4}$$

here $P2P(\cdot)$ denotes the ℓ_1 point-to-point distance and n_{DP} is the number of valid correspondences.

Temporal Pose Stability $\mathcal{L}^i_T$ is a temporal regularizer on SMPL poses. It is defined as the mean squared error of the current frame and the last frame SMPL joint rotation angles, which penalizes temporal pose jittering:

$$\mathcal{L}^i_T(\mathbf{P}_i) = \|\mathbf{P}_i - \mathbf{P}_{i-1}\|^2_2 \tag{5}$$

Pose Plausibility $\mathcal{L}^i_{\text{prior}}$, proposed in [3], reflects how plausible a pose is, given a pose prior learned from a large scale realistic pose corpus [1, 11]. The pose prior is modelled as a mixture of $N_{\text{gauss}} = 8$ Gaussian distributions with learned weights α_j, mean $\boldsymbol{\mu}_j$, and variance $\boldsymbol{\Sigma}_j$. The pose plausibility is given as:

$$\mathcal{L}^i_{\text{prior}}(\mathbf{P}_i) = -\log \sum_{j=1}^{N_{\text{gauss}}} \alpha_j \mathcal{N}(\mathbf{P}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{6}$$

2. Evaluation Protocol

2.1. Baselines

CAPE Following DSFN [5], we leverage the DSFN fitting pipeline to optimize the energy with respect to the latent codes of the CAPE model. Specifically, the publicly available checkpoints are taken and the objective is defined as latent codes' optimization for the CAPE decoder.

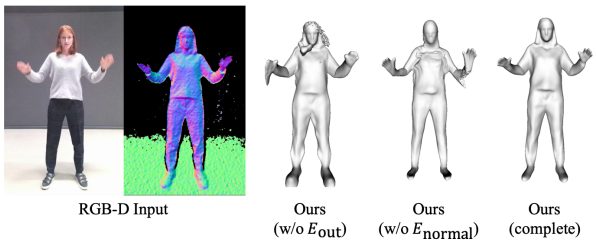

Figure 1. **Qualitative evaluation of energy functions.** Dropping the off-surface loss E_{off} results in many artifacts outside the human body. Ignoring the normal loss term E_{normal} leads to less detailed reconstructions.

SCANimate (2.5D) SCANimate [15] excludes concave regions from the smoothness constraint to avoid propagating incorrect skinning weights at self-intersecting regions. To adapt SCANimate [15] to our monocular depth setting, we slightly modify the concave region detection strategy. In SCANimate, concave regions are detected by computing the mean curvature on the surface of scans with a certain threshold. In contrast, we detect the concave regions by computing the point-to-surface distance between the point and the corresponding posed SMPL surface and remove the parts with a calculated distance larger than $4cm$.

3. Supplementary Result

3.1. Additional Ablation Studies

Effects of Energy Functions To gain insights into the optimization process, we perform an ablation study on the importance of the energy terms, which are introduced in Sec. 3.2. First, we drop the off-surface loss term E_{out} for the optimization. As shown in Fig. 1, this results in many artifacts outside the human body. This is because we only define losses on the body surface. Given partial point clouds of human bodies, the IGR loss alone is not sufficient to regularize the SDF. To tackle this problem, we leverage a loss term defined in Eq. 10 to regularize the off-surface SDF values. Another experiment is to compare our method with the baseline ignoring the normal loss term E_{normal} defined in Eq. 9. Shown in Fig. 1, this normal loss leads to more detailed reconstructions. Without this loss, the surface tends to be closer to the naked human body ignoring hair and cloth geometry.

3.2. Additional Qualitative Results

Our method PINA generalizes to various people with different human shapes and miscellaneous clothing styles. We show more qualitative results of our approach on real-world RGB-D sequences in Fig. 4 and in the **supplementary video**.

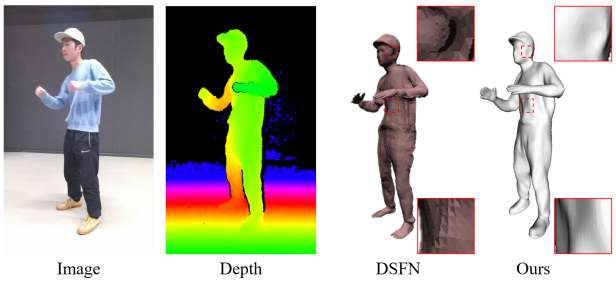

Figure 2. **Qualitative comparison on real data.** The reconstruction of DSFN can lead to holes and unrealistic surface noise. In contrast, our method ensures smooth surfaces while preserving fine surface details.

3.3. Comparison on real data

Due to the lack of accurate 3D ground-truth human scans for real monocular RGB-D inputs, we conduct qualitative comparisons to DSFN [5] on Azure Kinect sequences in terms of reconstruction accuracy. We show the result in Fig. 2. DSFN relies on the parametric human model SMPL and represents the clothing as a displacement layer on top of the minimally-clothed human body. This explicit representation strongly limits the reconstruction of surface details that geometrically differ from the minimally clothed human body and can cause holes (self-intersection) and unrealistic surface noise. In contrast, our method represents human shape implicitly and thus can better handle different topologies.

3.4. Comparison to fusion-based methods

We compare our method to BodyFusion [18] which is specifically designed to reconstruct dynamic human bodies from monocular RGB-D videos. The mean end-point error (EPE) is measured on the BodyFusion dataset. *Our* method attains an EPE score of **1.7 cm**, which significantly outperforms the specialized *BodyFusion* method (**2.2 cm**).

3.5. Comparison to SCANimate

To compare our method with current SOTA methods, we conduct quantitative and qualitative comparison with SCANimate. When comparing both methods in our setting, i.e., with 2.5D training data, our method (2.5D) outperforms SCANimate (2.5D) significantly as shown in Tab. 4 (in the paper) and in Tab. 1. We further compare both methods when using full 3D scans for training. Tab. 1 reveals that our method also outperforms SCANimate (3D) in all metrics. As shown in Fig. 3, our method can generate more realistic animation results for unseen poses compared to SCANimate. Overall, we show that our method outperforms SCANimate, irrespective of the setting.

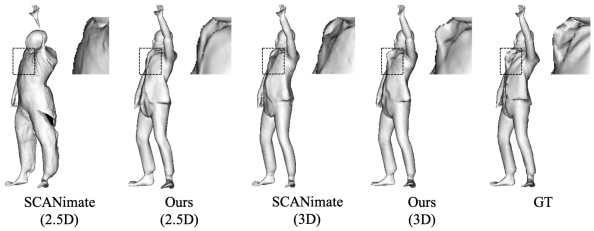

Figure 3. **Qualitative evaluation with SCANimate [15] on CAPE.** Our method produces more realistic results while animated compared with SCANimate, regardless of the type of input data.

Method	Input	IoU ↑	C $-\ell_2$ ↓	NC ↑
SCANimate [15]	2.5D	0.665	4.710	0.785
Ours	2.5D	**0.946**	**0.621**	**0.906**
SCANimate [15]	3D	0.941	0.560	0.906
Ours	3D	**0.949**	**0.501**	**0.920**

Table 1. **Quantitative evaluation with SCANimate [15] on CAPE.** Our method outperforms SCANimate regardless of the type of input data.

4. Limitations and Societal Impact Discussion

If regions of the clothed human are not visible in the input RGB-D sequence, the surface in these regions will show artifacts or be regularized to be smooth. Currently, our method does not model the appearance of the avatar. This will be an interesting direction for future work. Furthermore, our method performs well for garments that are topologically similar to the body. Non-skeletal induced dynamics are beyond the scope of this work. Combining our method with physics simulation to obtain a "personalized" garment simulation is an exciting direction for future work and can be leveraged to model and repose loose clothing like skirts.

PINA enables digitization of humans from a single commodity RGB-D sensor, which has many potential applications in movies, AR/VR, and telepresence applications. However, these may also lead to negative societal impacts, in particular privacy concerns such as deep-fakes. These must be addressed first before deploying virtual human avatars in products. Furthermore, our goal is clearly to enable positive uses of the technology developed here. While we cannot prevent nefarious uses, we argue that studying such technologies fully in the open, including discussion of technical details in the paper, release of code and data, should be preferred over closed, secretive study because it can also inform potential counter measures to deep-fake technology (if our work was to be abused for such purposes).

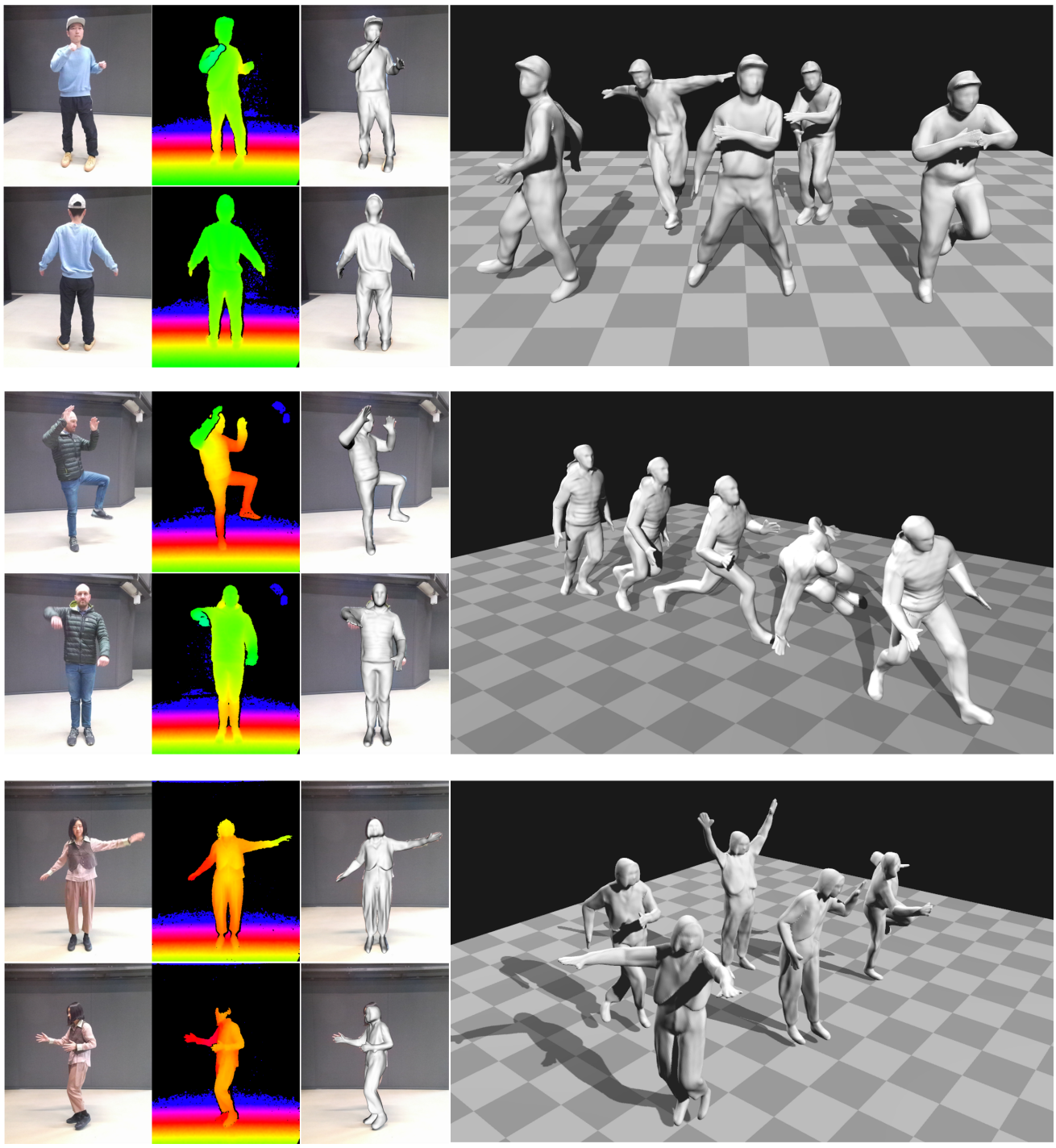

Figure 4. **RGB-D results.** We show qualitative results of our method based on real Azure Kinect RGB-D videos. Each subject has been recorded for 2-3 min (left). From noisy depth sequences, we learn shape and skinning weights (reconstruction), by jointly fitting the parameters of the shape and skinning network and the poses. We use motion sequences from [12, 13, 17] to animate the learned character.

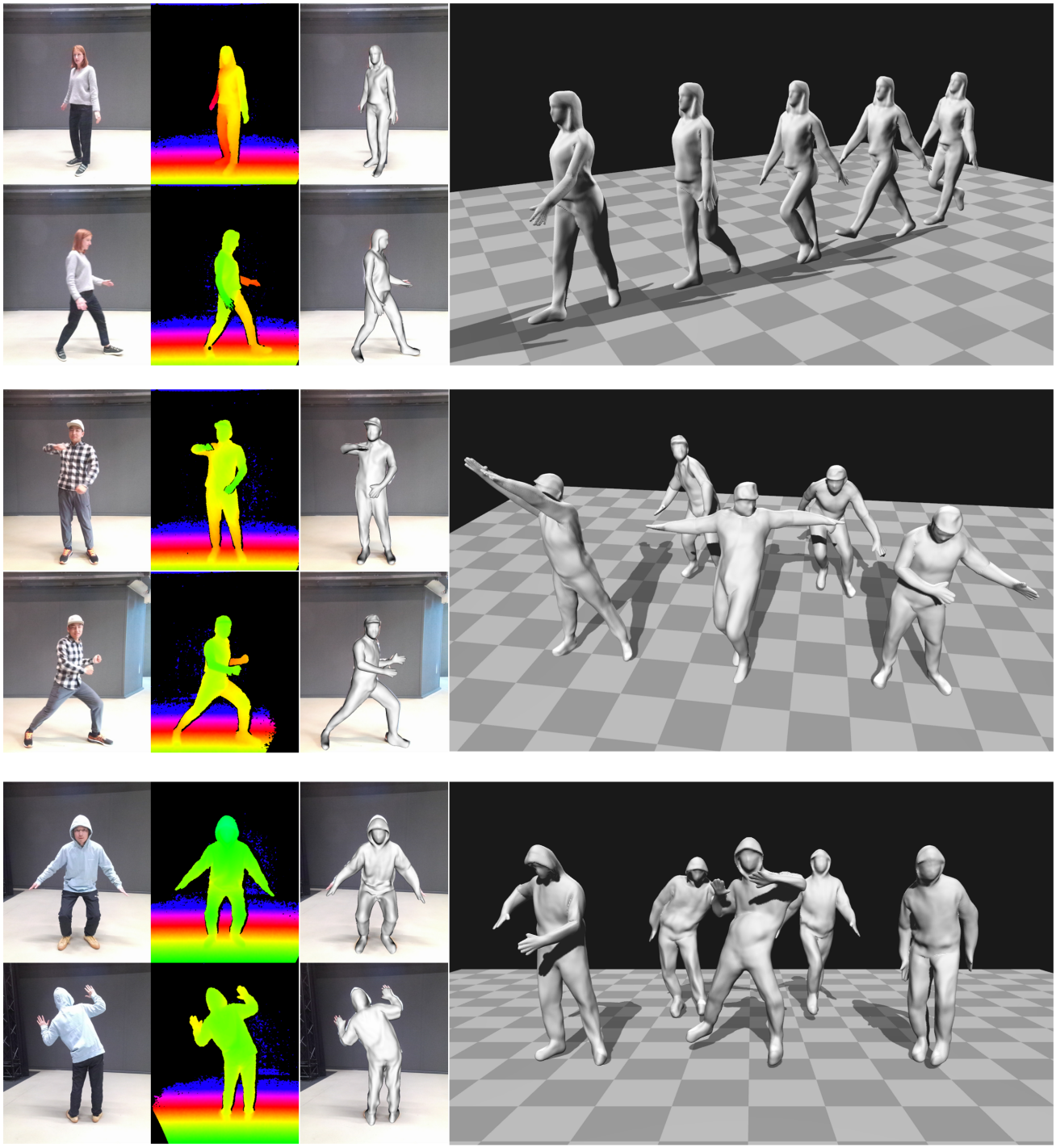

Figure 4. **RGB-D results.** We show qualitative results of our method based on real Azure Kinect RGB-D videos. Each subject has been recorded for 2-3 min (left). From noisy depth sequences, we learn shape and skinning weights (reconstruction), by jointly fitting the parameters of the shape and skinning network and the poses. We use motion sequences from [12, 13, 17] to animate the learned character.

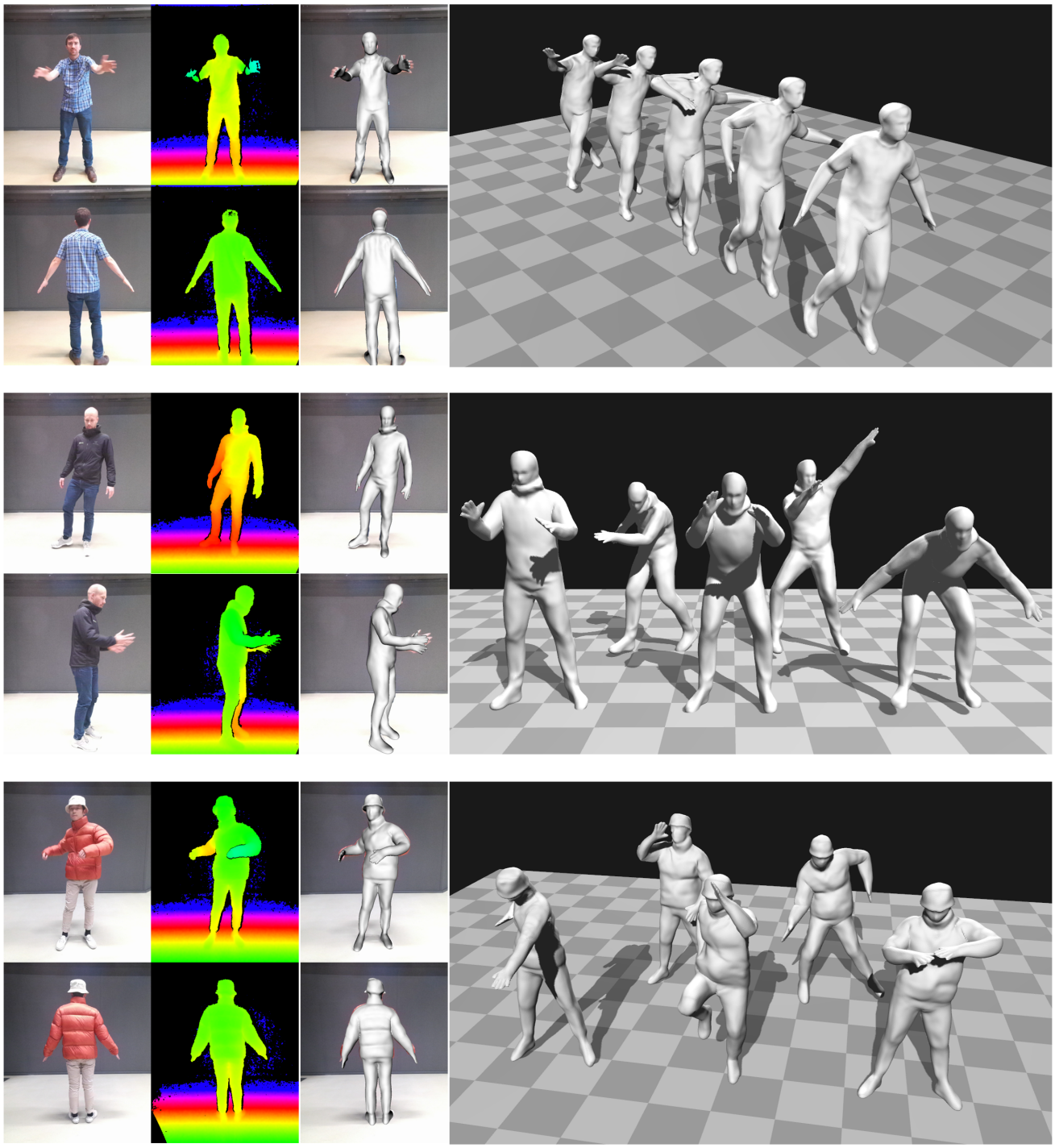

Figure 4. **RGB-D results.** We show qualitative results of our method based on real Azure Kinect RGB-D videos. Each subject has been recorded for 2-3 min (left). From noisy depth sequences, we learn shape and skinning weights (reconstruction), by jointly fitting the parameters of the shape and skinning network and the poses. We use motion sequences from [12, 13, 17] to animate the learned character.

References

[1] http://mocap.cs.cmu.edu. 2

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 1

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[4] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965. 1

[5] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021. 2, 3

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[7] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, pages 472–478, 2001. 1

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1

[11] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. 2

[12] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 5, 6

[13] 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, Oct. 2019. 4, 5, 6

[14] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

[15] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 1

[17] 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, Nov. 2019. 4, 5, 6

[18] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. 3