Dynamic Point Fields
Supplementary Material

Sergey Prokudin[1] Qianli Ma[1,2] Maxime Raafat[1] Julien Valentin[3] Siyu Tang[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems [3]Microsoft
[1] firstname.lastname@inf.ethz.ch

A. Extended Details on DPF Method

A.1. Normal Estimation and Transfer

Complementing main paper Section 3.1, here we elaborate on how to obtain point normals in the warped point clouds.

We consider three strategies to obtain a new point normal direction $\mathbf{n}_i^{(t)}$. First, in most considered scenarios, the points in canonical space represent vertices of a dense mesh, which are either given as ground truth or obtained by running a surface reconstruction algorithm [8]. Hence, after applying the transformation g_θ, we can keep the canonical space connectivity and directly obtain the new normal directions. If no such information is available, or the induced deformations lead to some drastic changes in mesh topology, we can also re-estimate normals using standard approaches [6]. Please note that all the steps above are fully differentiable, thanks to the modern deep learning 3D frameworks [7] and differentiable Poisson solvers [21].

Another way to obtain new normal directions is to directly warp them with the deformation network. We leave for future work an exploration of the correct architecture and parametrisation for such learning-based normal estimation, and provide only a proof-of-concept example of such approach in Figure A.1.

In general, thanks to the introduced isometric loss and a bias of the network towards smooth predictions, the predicted deformations in our experiments allow us to directly reuse the canonical space mesh topology; this strategy is implied, if not stated otherwise.

A.2. Optimisation Details

In most cases, we optimise the deformation field for 2000 steps, sampling 10^4 random points for computing $\mathcal{L}_{CD}$ and $\mathcal{L}_{iso}$. We use Adam optimiser [9] with an initial learning rate of 10^{-4}, decreasing the rate by 10^{-1} after each 200 steps of no improvement.

If a well-defined ground truth canonical surface is given (Sections 4.2-4.3 of the experiments), we omit the optimisation of point locations in the canonical space and optimise only the deformation field parameters w.r.t. $\mathcal{L}_{CD}, \mathcal{L}_{iso}, \mathcal{L}_V$.

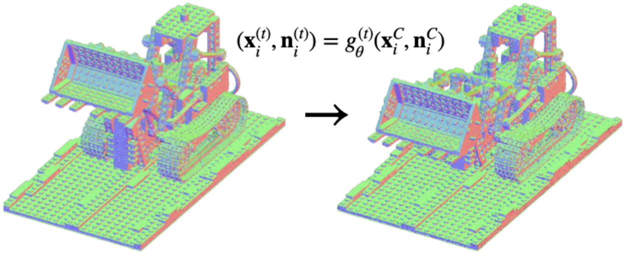
$(\mathbf{x}_i^{(t)}, \mathbf{n}_i^{(t)}) = g_\theta^{(t)}(\mathbf{x}_i^C, \mathbf{n}_i^C)$
→

Figure A.1. *Joint optimisation of the canonical space cloud and deformation network.* In this illustrative example, we showcase the simultaneous optimisation of the canonical cloud X^C and the deformation network g_θ. Apart from new point locations, the deformation network also predicts updated normal directions for each point. Please see the supplemental video for the animation of the sequence (title slide).

However, the possibility to optimise canonical space together with the deformation framework and to utilise rendering-based losses (e.g. $\mathcal{L}_{n_I}$) is essential for future work, when the introduced deformation paradigm will be combined with efficient photorealistic point-based renderers [29, 31]. We showcase feasibility of such optimisation in Figures A.1 and Figure 7 of the main manuscript.

B. Extended Experiments, Results, Discussions

B.1. Hyperparameter Selection, Implementation

For all the baselines considered in the paper, we aim to find the optimal parameters for optimisation and rendering. E.g., we experiment with the number of elements in the deformation pyramid [12] to achieve best results, find the best hyperparameters for the marching cube algorithm when comparing to SDF-based methods, etc.

We use original implementations of the NGLOD [25], NPMs [19], NDP [12] and Siren [24] frameworks [1–4]. We use the torch-NGP framework [26] for the instant NGP [17] baseline since it allows easy modification, and [3] for the Nerfies [20] and neural scene flow prior [10] baselines.

For the avatar animation baselines [13, 15, 22], we use the results kindly provided by the authors.

B.2. Point-based Surface Reconstruction

Other techniques. The proposed point optimization scheme can also be considered a hybrid of the differentiable surface splatting technique [30] and shapes-as-points approach [21]. Compared to DSS, we additionally utilize the Chamfer distance loss and a more efficient point renderer [28], which results in better reconstruction. Compared to SAP, our method has lower training time memory requirements since no differentiable Poisson solver needs to be ran on the grid for each pass. However, this technique can be complementary to our basic pipeline in case of the reconstruction from unoriented, noisy point clouds, as well as when a water-tight mesh reconstruction is required. The actual goal of this experiment is not to claim the ultimate advantage of any particular point optimization scheme, but rather show their appealing features compared to implicit methods.

B.3. Experiment: Learning Deformations

Additional results. Augmenting the discussion in Section 4.2 of the main manuscript, we provide additional comparisons with the best-performing non-rigid registration method [12] on the DeformingThings4D dataset in Figure B.3, as well as the failure case of our method (c).

SMPL-X guidance vs keypoint matching. To qualitatively justify the advantages of the SMPL-X based deformation learning guidance (Section 4.2), we visualise and discuss the results of the 3D keypoint matching algorithm [11] in Figure 4.

As-isometric-as-possible (AIAP) regularisation for SDF-based learning. We enforce the AIAP constraint on the points *sampled from the surface*. It is possible to apply the same kind of regularisation in the SDF case by simply enforcing it on *randomly sampled points in space*. However, the effect of this will be different. To illustrate this, we consider $\lambda_{iso} \to \infty$ (Eq. 19). Figure B.1 shows the results of this experiment on a simple scene consisting of two repelling 3D objects.

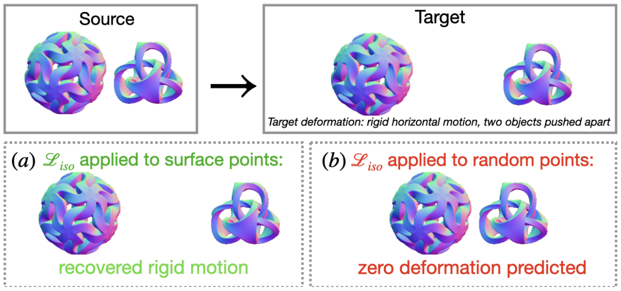
Source
Target
Target deformation: rigid horizontal motion, two objects pushed apart
(a) $\mathscr{L}_{iso}$ applied to surface points:
recovered rigid motion
(b) $\mathscr{L}_{iso}$ applied to random points:
zero deformation predicted

Figure B.1. *Learning a rigid horizontal motion of two 3D objects with* $\lambda_{iso} \to \infty$. Learning a deformation field for the scene while enforcing the AIAP loss only on *surface* points will result in recovering rigid motion of objects (a). In the case of enforcing same regularisation on *random* points in space (b), model will converge to predicting zero deformation everywhere. $\lambda_{iso} = 10^7$ in both cases.

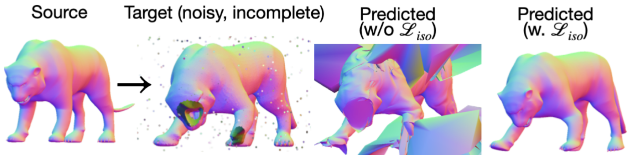
Source
Target (noisy, incomplete)
Predicted (w/o $\mathscr{L}_{iso}$)
Predicted (w. $\mathscr{L}_{iso}$)

Figure B.2. *Learning on noisy and incomplete scans*. Here we show the effect of learning a deformation field in the case of a target scan with noisy points (10^3) and holes (head, paw). While learning fails in the case of basic optimisation without AIAP loss (third column), utilising isometric loss leads to a plausible deformation (rigthmost row).

Experiments on noisy scans. While we generally consider learning on noisy or partial scans an avenue of future work, an example of learning a deformation for this case is showcased in Figure B.2.

B.4. Experiment: Avatar Animation

Here we provide extended details on the experiments in the main manuscript Section 4.3.

We conduct an experiment on the ReSynth [15] dataset. Each data frame t contains a dense point cloud $\{\mathbf{x}_i^{(t)}\}_{i=1,\cdots,N}$ ($N \approx 2 \times 10^5$ points) of a posed, clothed person , and its corresponding underlying minimally-clothed body mesh vertices $\{\mathbf{v}^{(t)}\}_{i=1,\cdots,N_v}$. The minimal body meshes have a consistent topology across all frames, but the clothed body point clouds do not have temporal correspondence. Given a set of such data pairs, we aim to produce clothed body for unseen test poses.

Implementation details. We adopt the guided deformation field learning described in the main manuscript Section 3.2. On a high level, we optimise a per-frame MLP to model the deformation field between minimal bodies of the canonical- and the target frame, and apply the optimised MLP to deform clothed body surface points.

Specifically, to predict a test frame t, we choose a canonical data frame $\mathcal{C}$ from the training set, and optimise a deformation field parameterized by an MLP $g_\theta^{(t)}$, such that the minimally-clothed body mesh vertices in the canonical frame $\{\mathbf{v}_i^{(\mathcal{C})}\}_{1,\cdots,N_v}$ matches that of the target frame $\{\mathbf{v}_i^{(t)}\}_{1,\cdots,N_v}$ after transformed by $g_\theta^{(t)}$:

$$\begin{aligned}\theta^* = \operatorname*{argmin}_\theta \sum_{i=1}^{N_v} \lambda_V ||\mathbf{v}_i^{(t)} - (\mathbf{v}_i^{(\mathcal{C})} + g_\theta(\mathbf{v}_i^{(\mathcal{C})}))||_1 \\ + \lambda_{iso}\mathcal{L}_{iso}.\end{aligned} \quad (1)$$

The optimisation is performed using the Adam optimiser with a learning rate of 1×10^{-4} and hyperparameters $\lambda_{iso} =$

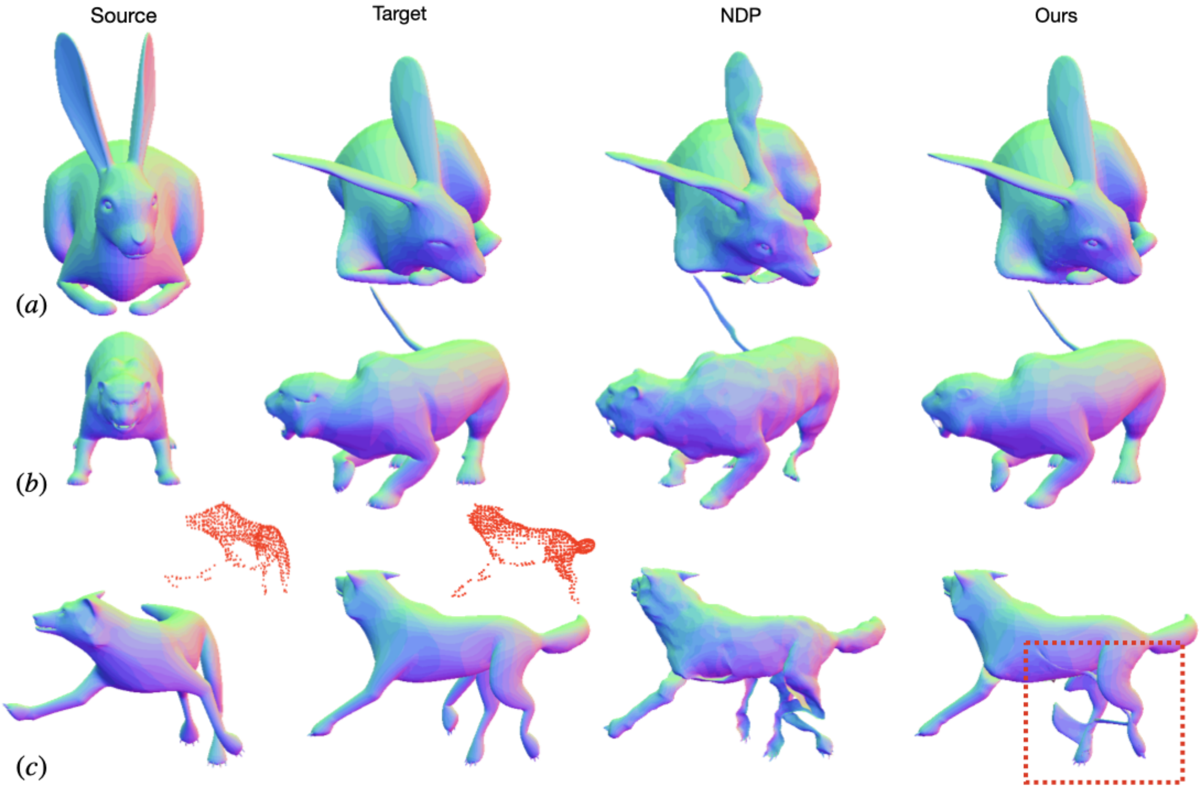

Figure B.3. *Additional results on non-rigid surface registration.* (a-b): comparison to the NDP [12] method in the case of Lepard [11] keypoint supervision. (c): failure case of our method in the situation of the highly articulated pose with missing keypoints. Matched points are visualised in red, pay attention to missing limbs.

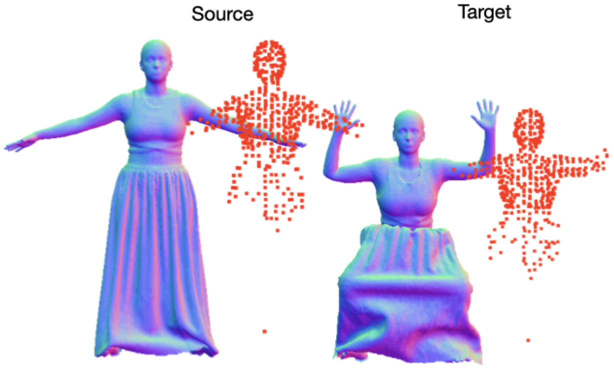

Figure B.4. *Detected Lepard [11] keypoints on the Resynth scan pair*. The off-the-shelf landmark detection method fails in the case of highly articulated humans and loose clothing, motivating the need for better deformation supervision.

$10^3, \lambda_V = 10^4$ for 2000 steps.

Once optimised, $g_\theta^{(t)}$ is used to transform arbitrary points on the canonical clothed body surface into the target pose. In practice, instead of deforming all points in the high-resolution source point cloud, we first perform Poisson Surface Reconstruction (PSR) [8] to obtain a mesh and deform the mesh vertices. In this way, we can transfer the mesh connectivities derived from PSR to the deformed point set and directly obtain the points' normals without the need to predict or re-estimate them, as discussed in Section A. Although the deformation field $g_\theta^{(t)}$ is optimised for the minimally-clothed body, the smooth inductive bias of the coordinate-based MLP results in a coherent clothing shape after deformation.

Choice of canonical frame. We investigate two strategies of choosing the canonical frame. (1) *"Single canonical frame"*: we simply choose the first frame (in an "A"-pose) from the ReSynth training set and use that as the source of deformation for all test poses. Empirically we find that the results from this approach are temporally coherent, as shown on the slide 16 of the supplemental video. It is also worth noting that this setting is essentially a solution to the "single scan animation" challenge [15], i.e. animating a given single scan to different novel poses. (2) *"Nearest canonical frame"*: for each test pose, we find its nearest pose (measured by the vertex-to-vertex distance of the underlying body mesh) in the training set, and use that frame as its canonical frame. This approach does not guarantee temporal consistency due to the discontinuity nature of the nearest neighbor search, but the produced clothing geometry varies more obviously with the body pose. We visualise several nearest canonical frames for various target poses in the Figure B.6.

Fairness of the comparison to baselines . Please note that the comparison with the baselines (SCANimate, POP, SkiRT) here is fair as we address exactly the same task using the same amount of information: all methods have access to a collection of scans and fitted SMPL bodies. The difference is in how we use them: the baseline methods use them as *training* data for a single feedforward model; our method uses them as a "pool" of source scans from which we retrieve a canonical scan at inference.

Discussions on the evaluation metric. In the main paper we compared with baseline methods using a perceptual study instead of the Chamfer distance and normal consistency metrics as used in recent papers [13, 15, 22]. Here we discuss the reason.

Previous works [13, 15, 22] evaluate clothing shape prediction accuracy by computing Chamfer and surface normal distances to the ground truth. This implicitly assumes a one-to-one mapping from body pose to the clothing shape but in reality, the clothing shape is not solely dependent on pose; it is also influenced by other factors such as the motion speed and history. Consequently, for a given pose, multiple clothing shapes can be plausible, as discussed in [5, 14]. An illustration is shown in Figure B.5.

Since the experiment in Section 4.3 is primarily purposed to demonstrate the representational power of our method and its applicability to human modeling, we resort to a perceptual study to characterize its quality, instead of adopting the pose-only regression metrics as discussed above.

Nevertheless, here we also provide the Chamfer distance and normal consistency errors for a reference in Table B.1. As our method deforms a source (canonical) cloud, the target state it produces, while plausibly looking, may not conform to that of the ground truth, hence the higher errors in the table. While pose-shape regression is not the focus of this work, we believe that combining it with our representation can lead to comparable to lower errors under these metrics while achieving higher visual quality as we show in the paper. We leave this for future work.

Method	anna-001		beatrice-025		christine-027		janett-025		felice-004	
	CD	NML	CD	NML	CD	NML	CD	NML	CD	NML
SCANimate [22]	1.34	1.35	0.74	1.33	3.21	1.66	2.81	1.59	20.79	2.94
PoP [15]	0.62	0.82	0.34	0.75	1.72	0.97	1.24	0.89	7.34	1.24
SkiRT [13]	0.58	0.81	0.31	0.77	1.54	0.99	1.10	0.82	6.45	1.25
Ours (single)	1.27	0.99	0.68	0.95	4.40	1.39	3.12	1.36	14.45	2.56
Ours (nearest)	0.96	0.96	0.46	0.99	2.88	1.24	2.51	1.19	16.07	2.50

Table B.1. Errors on pose-only regression metrics. Chamfer distance in $\times 10^{-4} m^2$; normal consistency in $\times 10^{-1}$. "Ours (single)": using a single frame as canonical frame; "Ours (nearest)": using nearest training scan per frame as the canonical frame (see Section B.4, *"Choice of canonical frame"*).

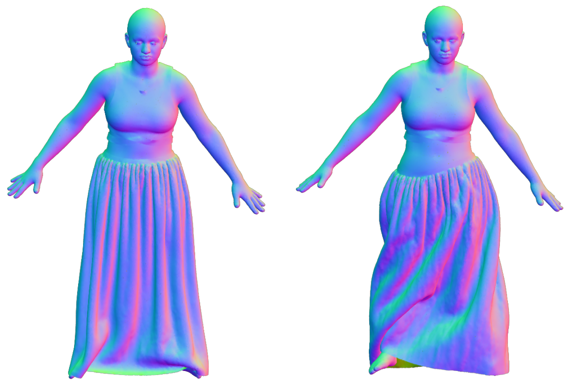

Figure B.5. *Stochastic clothing shape.* Given very similar body poses, the clothing shape state can largely differ, while both being valid. Example taken from the ReSynth [15] dataset, from the beginning and the end of the "hips" motion sequence, respectively. The body dynamics (motion history) has influenced the clothing shape despite the similarity in pose. The chamfer distance and normal consistency error between these two frames are $30.0 \times 10^{-4} m^2$ and 3.5×10^{-1}, respectively.

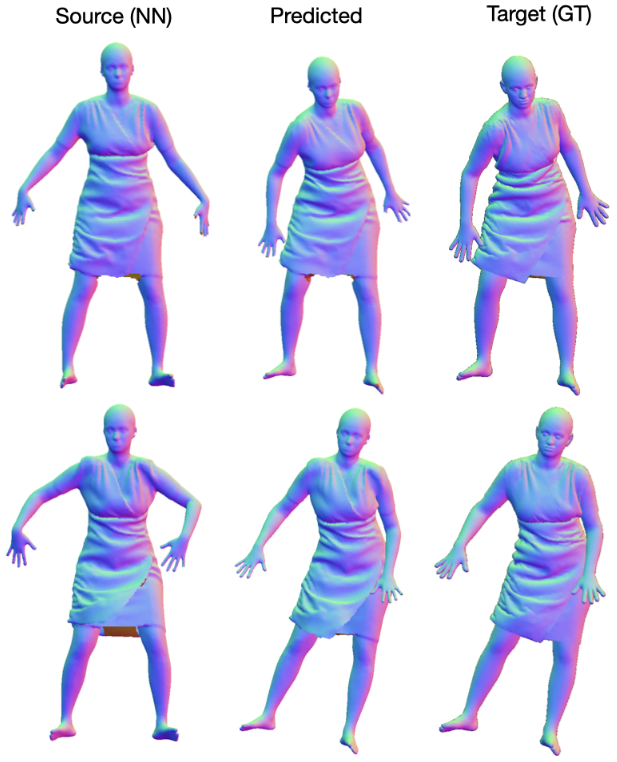

Figure B.6. *Nearest canonical frames.* Here, we visualise the nearest scans from the training set (left) used as a starting point for deformation modeling (middle), in the case of the "nearest scan" avatar animation regime. Please see the Section B.3 for more details.

Details on perceptual study. We choose the optimal renderering parameters for each method. For point-based baselines [13, 15], we use PyTorch3D point renderer with a point radius of 0.007 to minimize the visual artifact caused by the gaps between points. For SCANimate, we first extract a mesh from the generated implicit surface and render with the PyTorch3D mesh renderer. For our method we use the results generated by the *Nearest canonical frame* scheme, perform PSR (see paragraph *"Implementation details"*) and render the deformed mesh. The camera and lighting condition are kept same for all renderings.

The participants are presented with a set of 32 examples consisting of different subjects and poses, randomly sampled from the ReSynth dataset results. In each example, the ground truth rendering is shown on the left, and the results generated by different methods are presented in random order on the right, see Figure B.7. For each example, the participants are asked to select their most preferred, single result based on the following two criteria: (1) overall visual quality considering the realism of clothing shape, wrinkle details and level of artifacts; (2) resemblance to the ground truth. The statistics in the main paper Table 3 are computed by dividing the number of votes each method receives with the total number of votes (excluding 0.3% empty votes).

Result: extreme poses. Figure B.8 shows further qualitative results on predicting the clothing shape under extreme, out-of-distribution poses, in comparison to a recent point-based clothed human model, SkiRT [13]. Although SkiRT is designed to address the "split"-like artifacts for skirts and dresses by learning an LBS field for the garment, it inevitably reaches its limit under these poses: the points between the legs become very sparse. In contrast, our method well retains the completedness and smoothness of the cloth, showing a higher robustness for these garment types under extreme poses.

C. Limitations and Future Work

Our deformation optimisation pipeline still struggles to find a plausible deformation when large deformations are present and no guidance is available (Figure B.3*c*). This is especially pronounced in the case when the scene undergoes significant topological changes. One way to improve the deformation learning in the case of animals is to use estimators of the corresponding body joints [16, 18] and 3D models [32, 33].

Second, as discussed in the main paper, currently our method requires optimising a small network for each frame when dealing with unseen target scans or poses, which takes approximately 30*s*/frame with a NVIDIA RTX6000 GPU for the ReSynth data. This makes our method, in its current form, unsuitable for real-time applications. Potential solutions here include meta-learning approaches to learning deformations networks [23, 27].

Finally, when applied to modeling humans, the guided deformation optimisation does not explicitly learn the dependence of the clothing geometry on factors such as body pose and acceleration, which is an open research topic per se. Nevertheless, our experiments have demonstrated the capability of our approach in modeling highly non-rigid, articulated shape such as dynamic humans. Future work can combine our representation with domain-specific techniques to create high-quality real-time avatar models with explicit control on semantic and physical parameters.

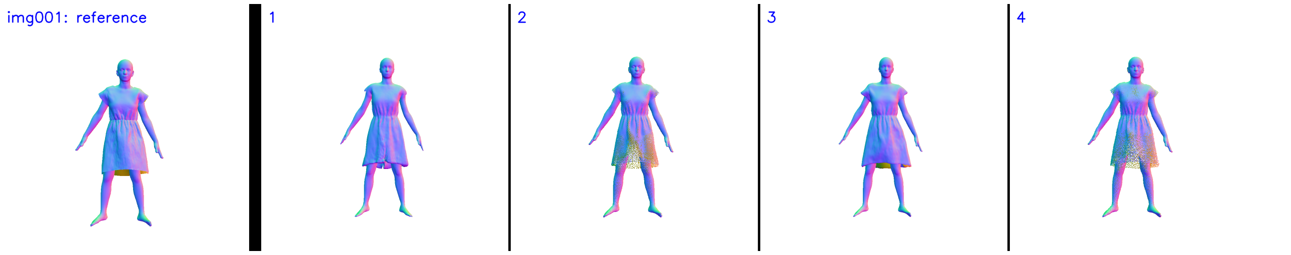

Figure B.7. *Example of perceptual study image.* The ground truth rendering is always presented on the leftmost position, and the ordering of results from different methods is shuffled per example.

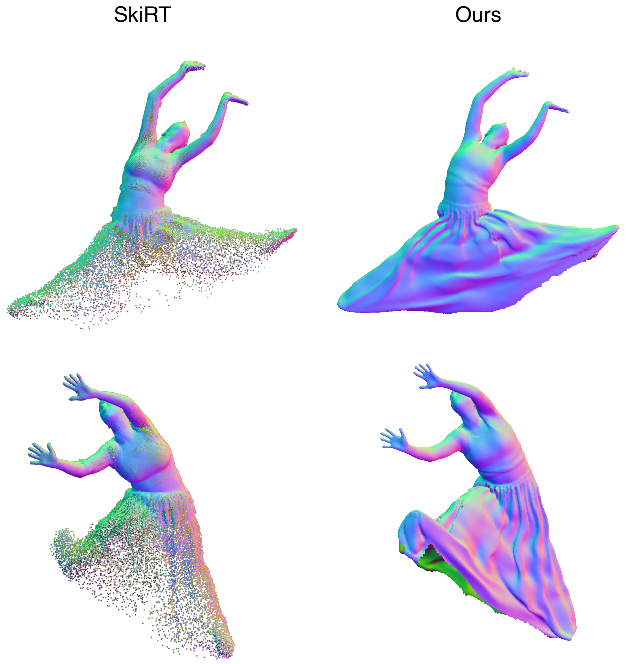

Figure B.8. Qualitative comparison against the SoTA point-based clothed human model [13] under extreme poses.

References

[1] https://github.com/nv-tlabs/nglod. 1

[2] https://github.com/pablopalafox/npms.

[3] https://github.com/rabbityl/deformationpyramid. 1

[4] https://github.com/vsitzmann/siren. 1

[5] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics*, 40(4):1–17, 2021. 4

[6] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Surface reconstruction from unorganized points. In *Proceedings of the 19th annual Conference on Computer Graphics and Interactive Techniques*, pages 71–78, 1992. 1

[7] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3D deep learning with PyTorch3D. In *SIGGRAPH Asia 2020 Courses*, pages 1–1. 2020. 1

[8] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3):1–13, 2013. 1, 3

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

[10] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:7838–7851, 2021. 1

[11] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[12] Yang Li and Tatsuya Harada. Non-rigid point cloud registration with neural deformation pyramid. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3

[13] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision (3DV)*, 2022. 2, 4, 5, 6

[14] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[15] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021. 2, 3, 4, 5

[16] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. 5

[17] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):102:1–102:15, July 2022. 1

[18] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using deeplabcut for 3D markerless pose estimation across species and behaviors. *Nature protocols*, 14(7):2152–2176, 2019. 5

[19] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural parametric models for 3D deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages

12695–12705, 2021. 1

[20] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[21] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[22] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2897, 2021. 2, 4

[23] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:10136–10147, 2020. 5

[24] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[25] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 1

[26] Jiaxiang Tang. Torch-ngp: a pytorch implementation of instant-ngp. https://github.com/ashawkey/torch-ngp. 1

[27] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 5

[28] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7467–7477, 2020. 2

[29] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-NeRF: Point-based neural radiance fields. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, 2022. 1

[30] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)*, 38(6), 2019. 2

[31] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–12, 2022. 1

[32] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 5

[33] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017. 5