Supplementary Materials for R-Cyclic Diffuser: Reductive and Cyclic Latent Diffusion for 3D Clothed Human Digitalization

Kennard Yanting Chan
Nanyang Technological University, Singapore
Institute for Infocomm Research, A*STAR
`kenn0042@e.ntu.edu.sg`

Fayao Liu
Institute for Infocomm Research, A*STAR
`liu_fayao@i2r.a-star.edu.sg`

Guosheng Lin
Nanyang Technological University, Singapore
`gslin@ntu.edu.sg`

Chuan Sheng Foo
Centre for Frontier AI Research, A*STAR
Institute for Infocomm Research, A*STAR
`foo_chuan_sheng@i2r.a-star.edu.sg`

Weisi Lin
Nanyang Technological University, Singapore
`wslin@ntu.edu.sg`

Table 1. Ablation on Blendweight SMPL-X Priors for Novel View Generation

	THuman2.0		
Methods	LPIPS ↓	PSNR ↑	SSIM ↑
w/o Blendweight SMPL-X Priors	0.0632	21.24	0.8933
w Blendweight SMPL-X Priors	**0.0373**	**24.69**	**0.9112**

1. Impact of SMPL-X Blendweight Priors on Novel View Generation

In our **main paper**, we demonstrated how our SMPL-X Blendweight Priors will influence our 3D reconstruction. In here, we will demonstrate how SMPL-X Blendweight Priors will affect our novel view generation (i.e. how SMPL-X Blendweight Priors affect the novel views generated by F).

Qualitatively, we can observe from Fig. 1 that when SMPL-X Blendweight Priors are used, the resulting novel views have a much more accurate body pose. Be informed that, as aforementioned in our **main paper**, the SMPL-X Blendweight Priors used in our experiments are derived from the groundtruth SMPL-X meshes.

Quantitatively, we compared the novel views generated by F when our SMPL-X Blendweight Priors are used and when these priors are not used. The results are shown in Tab. 1. From the table, we observed that when our SMPL-X Blendweight Priors are used, the novel views generated are closer to the groundtruth views compared to the case where these priors are not used.

2. Implementation of our Multi-view Pixel-aligned Implicit Model

In our R-Cyclic Diffuser, the final part of our pipeline consists of a multi-view pixel-aligned implicit model. Our implementation of this multi-view pixel-aligned implicit model is a modified version of the Multi-view PIFu outlined in [6].

Given N number of views, Multi-view PIFu would use a single-view PIFu for a total of N forward passes. In each forward pass, the single-view PIFu takes in a view and computes its corresponding feature embeddings. After the N forward passes are completed, the Multi-view PIFu will average up all the feature embeddings and compute a grid of SDF-like values. The grid is then used in the Marching Cubes algorithm to reconstruct a 3D mesh.

Our implementation also uses a single-view PIFu that does N forward passes. But instead of taking a simple average of the feature embeddings, we use a weighted average where the weights are dependent on the visibility of the corresponding views. For example, if a point is visible in view 1 but not visible in view 2, view 3, ... and view N, then the weight for the feature embedding for that point in view 1 will be 1.0 but 0.0 for the other views.

In addition, in each of the N forward passes, the single-view PIFu has access to the visual hull formed by the mask images of the N views. Concretely, the single-view PIFu has an additional 3D CNN module that takes in a voxelized visual hull and outputs voxel-aligned features that are combined with the other intermediate features produced by the

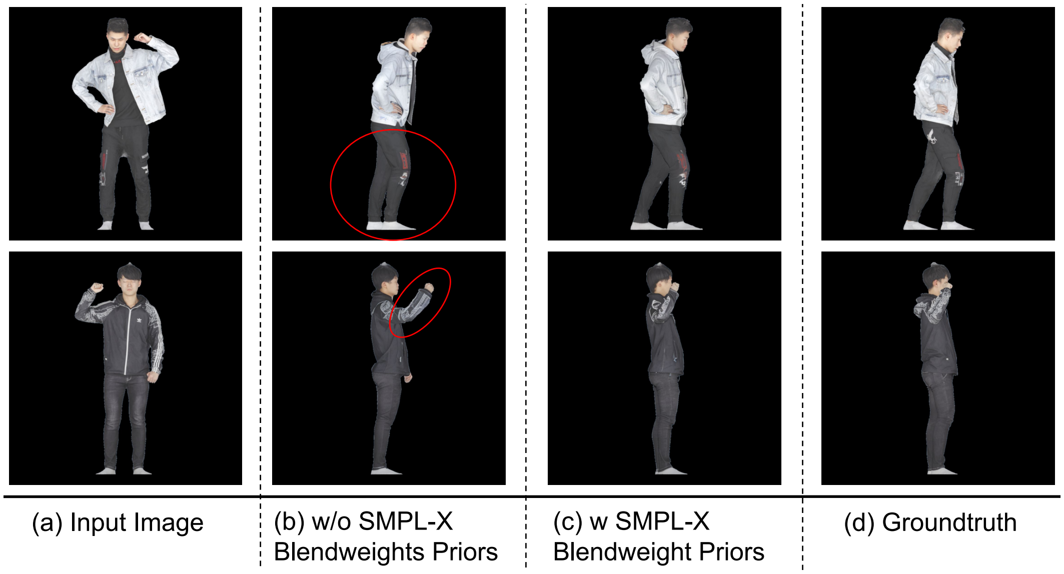

(a) Input Image
(b) w/o SMPL-X Blendweights Priors
(c) w SMPL-X Blendweight Priors
(d) Groundtruth

Figure 1. Ablation on Blendweight SMPL-X Priors for Novel View Generation

single-view PIFu. This is similar to how PaMIR [12] and DeepMultiCap [11] incorporate a SMPL parametric body model [2] into a PIFu.

If a SMPL-X mesh [5] is available, then we add an additional 3D CNN module to the single-view PIFu. This 3D CNN module would take in the SMPL-X mesh (in voxelized form) and output voxel-aligned features that would be used by the single-view PIFu. As aforementioned, this is similar to what has been done in [12] and [11].

After all the feature embeddings (produced in the N forward passes) are combined via weighted average and then used to compute a grid of SDF-like values, we use the Marching Cube algorithm [3] to obtain a 3D clothed human mesh. Lastly, we apply a refinement module used in ICON [8] to refine the 3D mesh. The refinement module uses normal maps that are derived from the RGB images predicted by our R-Cyclic Diffuser's F. Be informed that this refinement module is available in ICON's GitHub repository but not in their paper. To maintain a fair comparison, we use the same refinement module to refine the meshes produced by ICON when presenting ICON's results.

3. Qualitative Comparison of ECON and D-IF against our R-Cyclic Diffuser

In our **main paper**, we only compared ECON [9] and D-IF [10] against our R-Cyclic Diffuser quantitatively. Thus, we will now compare these existing models against our R-Cyclic Diffuser qualitatively. The results can be seen in Fig. 2. All models in the figure use the groundtruth SMPL-X meshes as inputs. The figure shows that ECON and D-IF struggle to reconstruct details in the sideviews and backviews of the 3D clothed meshes. In contrast, our R-Cyclic Diffuser is able to create very fine details on the sideviews and backviews of our meshes.

4. More Results on Internet Images

Here, we also show more results on real Internet images taken from Shutterstock. Please see Fig. 3. Similar to previously seen results, R-Cyclic Diffuser is able to reconstruct precise details on the front, back, and side views of meshes. Moreover, the figure shows that R-Cyclic Diffuser generalizes well to real images.

5. More Implementation and Hardware details

R-Cyclic Diffuser use the same pretrained Stable Diffusion network that has been used in Zero-1-to-3 [1]. This Stable Diffusion network is pretrained on the LAION datasets [7].

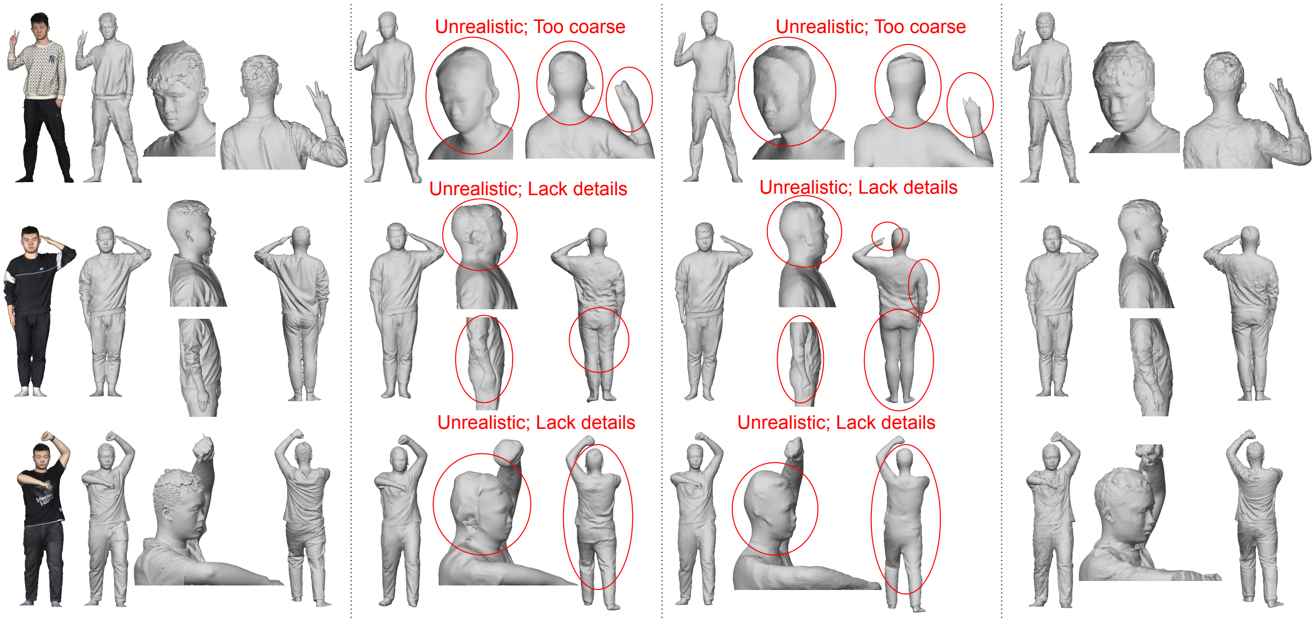
As mentioned in [1], this pretrained Stable Diffusion network is specially tweaked such that it can accept conditional information from images.

During finetuning with THuman2.0 images, we use an AdamW [4] optimizer with a learning rate of 10^{-4}. We use a batch size of 72, an image resolution of 512×512, and a latent dimension of $64\times64\times4$.

We finetuned our R-Cyclic Diffuser using NVIDIA RTX A5000 GPUs.

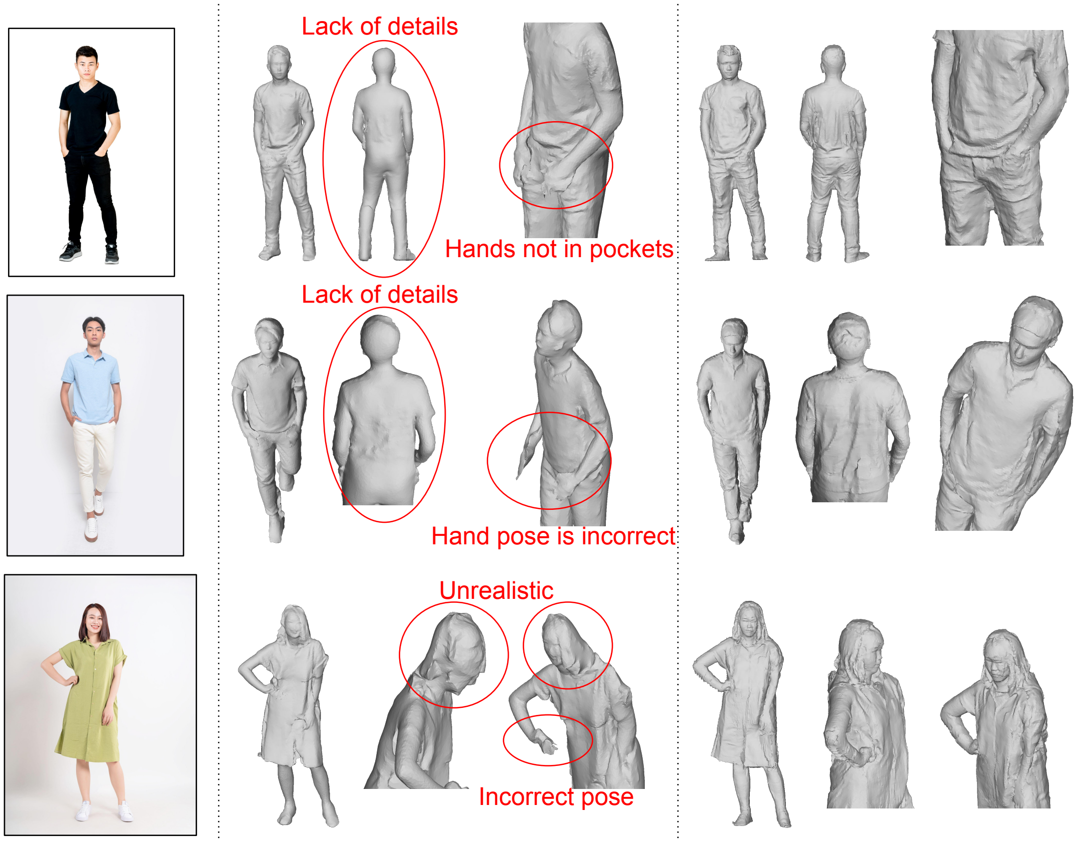
During inference, we set the number of DDIM sampling steps to 200 except during Cyclic Noise Selection, where we reduce the number of steps to 40 in order to identify the best set of Gaussian noises.

More implementation details can be found in our source code, which will be publicly released.


Unrealistic; Too coarse
Unrealistic; Too coarse
Unrealistic; Lack details
Unrealistic; Lack details
Unrealistic; Lack details
Unrealistic; Lack details

(a) Groundtruth (b) ECON (c) D-IF (d) Ours w SMPL-X

Figure 2. Qualitative Comparison of ECON [9] and D-IF [10] against our R-Cyclic Diffuser.



(a) Input Image (b) ICON (c) Ours w/o SMPL-X

Figure 3. More results on real images retrieved from Shutterstock.

References

[1] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2

[2] 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. 2

[3] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[5] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[6] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1

[7] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[8] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[9] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 2, 3

[10] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9122–9132, 2023. 2, 3

[11] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 2

[12] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 2