Methods for Human Digital Avatar Reconstruction

Ostap Viniavskyi[1]

DressX
ostap.viniavskyi@gmail.com

Abstract. The demand for high-fidelity human avatar reconstruction rapidly increases across various industries, including e-commerce, gaming, and virtual reality. Current methods often rely on complex motion capture systems or extensive multi-view video data, making them impractical for widespread use. This proposal describes the research that aims to develop a novel approach to 3D human digitalization that overcomes these limitations by enabling the creation of detailed, animatable, and customizable digital avatars from minimal input, such as a single video captured by a smartphone.
The project will focus on several key challenges in the field: disentangling the human body and clothing representations, accurately modeling loose clothing, and optimizing both the quality and efficiency of the reconstruction process. By leveraging recent advancements in 3D reconstruction techniques and neural networks, the proposed method seeks to achieve industry-grade results while minimizing the need for specialized hardware and complex data collection processes.
The outcomes of this research have the potential to significantly impact the way digital humans are created and utilized, offering more accessible and scalable solutions for industries that require realistic and interactive avatars. The successful implementation of this work will not only advance the current state of the art but also contribute valuable insights and tools to the broader field of human digitalization.

Keywords: 3D Avatar · 3D Human Digitalization · Human Understanding · Novel-view synthesis · Implicit Representations · Computer Vision (CV)

1 Introduction

Human avatar reconstruction or human digitalization is a task that involves creating a digital twin of the person from single- or multi-view images or video data [12]. Ideally, the avatar should preserve as many details of the human body as possible, such as correct shape estimates and detailed clothing reconstruction, as well as preserve human identity.

Another important aspect of the digital avatar is its ability to interact with the digital world, where it will be placed. The avatar should be animatable, meaning it can be placed in any feasible pose that might differ from the poses seen

at the time of capture. Also, the reconstructed avatar should correctly interact with the lighting of a new environment. In more complex scenarios, the digital twin has to be customizable, which stands for the ability to change different avatar parts, such as swapping the clothes, changing body shape, hairstyle, etc., or even combining different aspects from multiple reconstructions [5].

Finally, we aim to make the reconstruction process as lightweight as possible, without the need for a complex 3D capture system, while the rendering of the reconstructed digital avatar should be in real-time [5].

The importance of addressing these challenges is underscored by the potential benefits that high-fidelity human avatars can bring to various industries. In the entertainment sector, for example, realistic digital humans can significantly enhance the immersion of virtual reality experiences and gaming environments. In e-commerce, accurate digital avatars can enable more personalized shopping experiences by allowing customers to try on clothing virtually. Additionally, developing animatable human avatars is crucial for advancing the Metaverse [19] and holoportation technologies [21], where real-time interaction in virtual spaces is becoming increasingly important.

Modern approaches to industry-level digital human avatar reconstruction usually rely on complex motion capture systems that use a combination of wearable devices, such as Inertial Measurement Units, multiple synchronized cameras, and calibrated lighting settings to capture the fine-grained details of the human bodies. Yet, complicated post-processing is still required on the artists' side, requiring tons of manual work.

Recently, developments in both implicit and explicit methods for 3D reconstruction and novel view-synthesis, such as NeRF [20], 3D Gaussian Splattings (3DGS)[10], and 3D human body representation [15], allowed significant progress in the field of human digitalization. Multiple approaches to 3D human avatar reconstruction achieve promising results using only monocular video from common devices such as a single digital camera or, with the help of generative models [8], can successfully reconstruct human bodies even from a single image. Yet many unsolved problems in the field require rigorous research. The more detailed state of the field and the challenges that are currently faced are given in the next section.

2 Related Works

3D mesh is among the most popular representations of 3D objects, particularly human avatars. The early approaches [25], [6] tried to reconstruct a textured mesh of a human subject and then use physical simulation to animate it. Additionally, more recent approaches [2], [27] utilized neural networks to implicitly learn the textures and motion of human bodies. The methods are complex in their nature as they usually require tracking non-rigid deformations of human bodies and cloth across multi-view videos. Additionally, one must pre-scan the subject's template to obtain a canonical representation.

With the rise of implicit models, many NeRF and SDF-based methods were developed for human avatar modeling. Animatable NeRF work [22] proposed to add SMPL deformations to the NeRF for modeling rigid motions of the human bodies. Neural Actor [13] parametrizes the human body on the surface of the SMPL avatar yet cannot model humans wearing loose clothes far from the template SMPL body. Other methods [3], [24] model different body parts of humans separately, such as body, face, and hands, learning separate implicit functions for each part. Nevertheless, implicit representation-based methods cannot achieve high-fidelity results. This is caused by the low-frequency bias that is present in the MLP [26], which is the foundation of the implicit methods. Additionally, the rendering of implicit models takes a long time because of the procedure that makes multiple samples along the ray. Special acceleration methods are needed to achieve at least near-real-time rendering speed.

Points-based methods are another promising direction for animatable human avatar representation. [18], [17] learn non-rigid pose-dependent deformations of the human body in the SMPL UV space, which helps modeling complex cloth motion. However, these methods require a 3D scan of the human body in advance.

More recent approaches to human avatar modeling utilize another points-based representation, 3D Gaussian Splattings. 3DGS [10] showed their effectiveness in representing and fast real-time rendering of complex 3D scenes. 3DGS methods for avatar modeling usually start from the template mesh of the human body, points on which serve as initialization of the splats. 3DGS-Avatar [23] learn the splats in the canonical space defined around the SMPL avatar in the rest pose. The authors use the learnable Linear Blend Skinning function for rigid deformations and the neural network for learning non-rigid pose-dependent deformations to achieve impressive results in human reconstruction and generalization to unseen poses. When rendering novel views, the rays cast from the camera are transformed into the canonical space, allowing for learning the splat's viewing-direction-dependent color even from the monocular camera. The main downside of the method is the choice of the template avatar as SMPL. Since most splats are initialized on the surface of the SMPL body, leaning loose clothes remains challenging for this approach.

In contrast to the previous method, [11] creates a subject-dependent template mesh, which allows for the modeling of wide clothes. The method starts with synchronized multiple-view frames from several videos, which capture the human in the pose that is closest to the rest pose. The template mesh is extracted from them through the SDF fitting procedure, and then Marching Cubes [16]. The mesh is assigned Linear Blend Skinning (LBS) weights, which are diffused from the SMPL. 2D Convolutional neural networks are used to directly predict Gaussian Splattings in each pixel of the front and back renders of the template mesh. Additionally, prediction materials properties and utilizing a differentiable Physically-Based renderer allow the creation of a relightable human avatar. The downside of the approach is the need for multiple-view video data, which is not easy to obtain in a real-world scenario.

In addition, most of the existing methods focus on the composite representation of clothed humans, which doesn't allow easy disentanglement of the body and cloth representations. Also, modeling the components not driven by the underlying human skeleton, such as hair, remains challenging.

3 Motivation

I selected this topic for research because of its importance in multiple industries like e-commerce, game development, and virtual reality-enhanced content creation. Creating high-quality avatars that can be reconstructed by off-the-shelf means will benefit these industries by saving up lots of manual work of 3D artists, spending on expensive equipment will create more immersive augmented and virtual reality effects.

Based on the field's current state, which was reviewed in the previous chapter, we can conclude that many open problems remain. Among them, we highlight the most prominent for future research.

Disentanglement of the human avatar representation. 3D reconstructions of human bodies mostly treat the human body and clothes together without the possibility of trivially disentangling them. Extracting separate representations for the clothes and body is a challenging task because of the occlusions and complex dynamics of cloth, body, and environment interaction. How to robustly extract cloth and its dynamical and optical properties remains an open question.

Working with loose clothes. Most of the existing approaches work with the human body template unsuitable for modeling loose clothes. Those that dynamically predict templates require multiple-view input or achieve low quality because of the utilization of generative models. Reconstruction of human bodies with loose clothes, specifically from single-view input, remains an open challenge.

Quality and speed of reconstruction. Even though significant progress has been made in this part in recent years, the quality of 3D human reconstruction is still not comparable with industry-grade expensive motion capture systems. The results produced by the approaches are based on implicit representations, are over-smoothed, and do not capture small details about humans and clothes. Much more progress should be made in improving the output reconstruction resolution, training, and rendering time.

Reconstructing from sparse views The final result of 3D human digitalization depends greatly on the quality and density of the input data. State-of-the-art approaches rely on the multi-view video input that is sometimes unrealistic to capture in real-world cases. How to achieve high-quality reconstruction from minimal inputs is an unsolved problem. More extreme scenarios might even include reconstruction from a single image, which must include a significant part of generative modeling to imagine the unseen parts of the reconstructed humans.

In addition, this research direction greatly aligns with my background and my future career goals:

- A significant part of my previous experience includes working with human understanding tasks like human parsing, DensePose [7] estimation, 3D human pose, and shape estimation. I have experience working with multiple human representations in 2 and 3 dimensions and those that make a bridge between them.
- The central part of my Master's thesis was working with 3D representations of humans and how they can be utilized for synthetic human dataset generation.
- My work experience in DressX company includes working with Virtual Try-On technologies. I explored generative and rendering approaches to the try-on task in depth and worked with different cloth representations, starting from ordinary images and finishing with 3D models that the physically-based engine can simulate. From my working experience, I gained valuable insights into the need for robust and efficient 3D human and cloth digitalization technologies, which can be further used in the try-on engines deployed by e-commerce companies.
- I believe that the work on 3D human digitalization is a logical continuation of my previous experience and a reasonable step forward, considering my prior knowledge of the domain. The successful implementation of the solutions to the proposed problems will allow me not only to publish scientific papers on the matter but also directly boost the industry via my work.

4 Approach

As described previously, 3d human digitalization is a problem that is far from being solved with the multiple aspects that need refinement, one of which is the efficiency of the algorithms during reconstruction and rendering times. The natural approach to solving such a multi-dimensional problem would be to first aim for maximum quality of the results by utilizing all possible data and resources and then optimizing for efficiency.

Firstly, we will need to establish a baseline method and explore available data and metrics on which the algorithms are trained and evaluated. For the baseline, the potential candidates are 3DGS-based methods like [23] and [11], which offer state-of-the-art avatar reconstruction form single- and multiple-view videos respectively.

Data collection: There exist a bunch of available datasets for the fitting and evaluation of avatar estimation tasks. THuman4.0 [28], AvatarReX [29], ActorsHQ [9] dataset provide multiple-view sequences ideal for fitting avatar models. Usually, a separate model is learned for each data sequence, with some frames left for evaluation. Additionally, monocular datasets like ZJU-MoCap [22] and PeopleSnapshot [1] can be used for a similar purpose.

Evaluation protocols: The evaluation is usually performed both qualitatively and quantitatively. For the quantitative part, reconstruction metrics are computed in the image space on the frames, which are held out for validation and testing in the framework of novel view synthesis. The common metrics include

Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Frechet Inception Distance (FID).

Template mesh construction: An important aspect of all modern approaches is the choice of template mesh. Monocular methods usually rely on the default template mesh, choosing SMPL for this task. Multi-view methods can utilize multiple frames at a particular time instance and try to reconstruct the template surface from them. Since we aim to minimize input data sources, we will proceed with the monocular setting, at the same time reconstructing a subject-dependent template mesh that can include loose clothes. In such cases, we will need to rely on either a single frame for reconstruction or multiple unsynchronized frames from different time steps. A potential solution may include the ideas from zero-shot single-image 3D object generation approaches like [14].

Body and cloth disentanglement: The industry-grade system for 3D human digitalization will require the disentangled representation of the human body and cloth. This will require accurate segmentation, potentially in 3D, of the clothes from the human body. Additionally, one will need to learn the deformations of the clothes that arise from the body's pose changes in order to achieve cloth transfer between multiple subjects. An approach similar to SCARF [4] might be taken where separate implicit models are learned for body and cloth based on the 2D segmentation.

Use-cases and integration: Finally, the approach should be efficient in terms of data and computational resources. We aim to create an approach that can run from a single video captured by a smartphone camera. Ideally, the learning stage should take up to a few minutes, with the rendering possible on the mobile device.

5 Summary

The proposed research aims to address critical challenges in the field of human avatar reconstruction by developing methods that enable high-fidelity, animatable, and customizable digital humans from minimal input data. By focusing on the disentanglement of body and cloth representations, handling loose clothing, and optimizing both the quality and efficiency of reconstruction processes, this project seeks to advance the state-of-the-art in 3D human digitalization.

Successful completion of this research will not only contribute significantly to academic knowledge in the domain but will also have practical applications in industries such as e-commerce, gaming, and virtual reality. By making the avatar reconstruction process more accessible and efficient, this work has the potential to revolutionize the way digital humans are created and utilized, aligning with both the research goals and the future career aspirations of the candidate.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models (2018), https://arxiv.org/abs/1803.04758

2. Bagautdinov, T., Wu, C., Simon, T., Prada, F., Shiratori, T., Wei, S.E., Xu, W., Sheikh, Y., Saragih, J.: Driving-signal aware full-body avatars. ACM Transactions on Graphics **40**(4), 1–17 (Jul 2021). https://doi.org/10.1145/3450626.3459850, http://dx.doi.org/10.1145/3450626.3459850
3. Dong, J., Fang, Q., Guo, Y., Peng, S., Shuai, Q., Zhou, X., Bao, H.: Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. In: Advances in Neural Information Processing Systems (2022)
4. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video (2022), https://arxiv.org/abs/2210.01868
5. Gleicher, M.: Animation from observation: Motion capture and motion editing. SIGGRAPH Comput. Graph. **33**(4), 51–54 (nov 1999). https://doi.org/10.1145/345370.345409, https://doi.org/10.1145/345370.345409
6. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: Drape: Dressing any person. ACM Trans. Graph. **31**(4) (jul 2012). https://doi.org/10.1145/2185520.2185531, https://doi.org/10.1145/2185520.2185531
7. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild (2018), https://arxiv.org/abs/1802.00434
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), https://arxiv.org/abs/2006.11239
9. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. ACM Transactions on Graphics **42**(4), 1–12 (Jul 2023). https://doi.org/10.1145/3592415, http://dx.doi.org/10.1145/3592415
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering (2023), https://arxiv.org/abs/2308.04079
11. Li, Z., Sun, Y., Zheng, Z., Wang, L., Zhang, S., Liu, Y.: Animatable and relightable gaussians for high-fidelity human avatar modeling (2024), https://arxiv.org/abs/2311.16096
12. Lin, Y., Chen, L., Ali, A., Nugent, C., Ian, C., Li, R., Gao, D., Wang, H., Wang, Y., Ning, H.: Human digital twin: A survey (2022), https://arxiv.org/abs/2212.05937
13. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control (2022), https://arxiv.org/abs/2106.02019
14. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023), https://arxiv.org/abs/2303.11328
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
16. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. p. 163–169. SIGGRAPH '87, Association for Computing Machinery, New York, NY, USA (1987). https://doi.org/10.1145/37401.37422, https://doi.org/10.1145/37401.37422
17. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: Scale: Modeling clothed humans with a surface codec of articulated local elements (2021), https://arxiv.org/abs/2104.07660
18. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing (2021), https://arxiv.org/abs/2109.01137

19. Meepung, T., Kannikar, P.: Metaverse; virtual world challenges and opportunities for digital business. Journal of Economics, Business and Management **10**, 277–282 (01 2022). https://doi.org/10.18178/joebm.2022.10.4.712
20. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020), https://arxiv.org/abs/2003.08934
21. Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P.L., Khamis, S., Dou, M., Tankovich, V., Loop, C., Cai, Q., Chou, P.A., Mennicken, S., Valentin, J., Pradeep, V., Wang, S., Kang, S.B., Kohli, P., Lutchyn, Y., Keskin, C., Izadi, S.: Holoportation: Virtual 3d teleportation in real-time. p. 741–754. UIST '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2984511.2984517, https://doi.org/10.1145/2984511.2984517
22. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies (2021), https://arxiv.org/abs/2105.02872
23. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting (2024), https://arxiv.org/abs/2312.09228
24. Shen, K., Guo, C., Kaufmann, M., Zarate, J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars (2023)
25. Stoll, C., Gall, J., de Aguiar, E., Thrun, S., Theobalt, C.: Video-based reconstruction of animatable human characters. ACM Trans. Graph. **29**(6) (dec 2010). https://doi.org/10.1145/1882261.1866161, https://doi.org/10.1145/1882261.1866161
26. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains (2020), https://arxiv.org/abs/2006.10739
27. Xiang, D., Bagautdinov, T., Stuyck, T., Prada, F., Romero, J., Xu, W., Saito, S., Guo, J., Smith, B., Shiratori, T., Sheikh, Y., Hodgins, J., Wu, C.: Dressing avatars: Deep photorealistic appearance for physically simulated clothing. ACM Transactions on Graphics **41**(6), 1–15 (Nov 2022). https://doi.org/10.1145/3550454.3555456, http://dx.doi.org/10.1145/3550454.3555456
28. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling (2022), https://arxiv.org/abs/2203.14478
29. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: Avatarrex: Real-time expressive full-body avatars (2023), https://arxiv.org/abs/2305.04789