

From a Single Image to a 3D Model: A Review of 3D Human Reconstruction from Monocular Images

Bolin Guan

*Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China
1250028470@student.must.edu.mo*

Abstract. In recent years, 3D human reconstruction technology has been widely applied across various industries such as digital humans, virtual reality, gaming, and film production. Among them, the task of generating a 3D human body model from a single two-dimensional image, known as single-image 3D human body reconstruction—poses significant technical challenges due to the relatively lack of input information. This article systematically reviews the mainstream technical routes and their advantages and drawbacks from 2015 to 2026. It starts from the early method of Skinned Multi-Person Linear Model (SMPL) parametric template, then progresses to the voxel prediction method represented by Deephuman, followed by the implicit function method centered on the Pixel-Aligned Implicit Function (PIFu) series as the benchmark, and finally concludes with the emerging neural rendering and real-time reconstruction technologies in recent years. The three significant experimental datasets, namely PIFu, Pixel-Aligned Implicit Function High-Resolution (PIFuHD) and Geometry and Pixel-Aligned Implicit Function (Geo-PIFu), which have milestone importance, are analyzed using indicators such as chamfer distance, P2S and normal error. The improvement logic and applicable boundaries of each method are also discussed. Finally, the current practical challenges faced in this field, such as data deficiencies and insufficient social popularity, are summarized. The future technologies, like large model frontier inference and real-time rendering, are also prospectively discussed.

Keywords: Three-dimensional human body reconstruction, implicit function, three-dimensional radiation field, 3D Gaussian Splatting

1. Introduction

It's well known that with the extensive development of character modeling in games, high-precision digital doubles play a more and more significant role in film and television special effects. The deep promotion of virtual reality and virtual idols promotes the market demand for 3D human body model reconstruction rapidly. The acquisition of traditional 3D data relies on specialized equipment such as multi-camera arrays and depth scanners, which are extremely expensive and difficult to operate and thus not widely available. Photos merely record 2D information. Its depth in imaging has already been lost. As a result, transitioning from 2D to 3D regionally is the issue of incomplete information.

In the past 10 years, scientists have explored plenty technical approaches. The earliest method was based on parametric template design. For example, SMPL model controlled human body deformation by using shape and pose parameters [1]. This kind method generally had stable results. But they were essentially nude models and could not reconstruct details like clothes, hair and face features. The detail processing was weak and far from authenticity.

Around 2019, two pioneering works broke through the limitations in this field. DeepHuman was the first to successfully integrate neural networks into model reconstruction. It creates a representation of the human body by voxels, allowing the network to directly predict the voxel grid from images. Divide the three-dimensional human body into countless tiny units and reorganize them in a reasonable manner. The advantage is intuitive, but the drawback is high computational load, low resolution [2]. Almost at the same time, PIFu brought a completely different technical route. It no longer used voxels, but instead used implicit functions to represent the human body [3]. Because implicit functions were not limited by resolution. PIFu was the first to reconstruct a clothed human body with wrinkles, directly addressing the limitation of DeepHuman's insufficient resolution. However, PIFu also had its own problems. That was not stable enough in complex postures (such as bending over, arms crossed), and prone to shape distortion. Later, PIFuHD increased the input image resolution from 512 to 1024, making the details of fingers and facial features more clearly [4]. While Geo-PIFu added global geometric constraints to the network, specifically solving the stability issues in complex postures [5]. Both works targeted the shortcomings of PIFu, but they have different focuses. One addressed the input level, and the other did the network structure level.

By 2022, Implicit Clothed humans Obtained from Normals (ICON) found a different approach and made a big difference. It brought SMPL back, but used it in a different way. ICON, used SMPL as "prior knowledge", first building a basic framework, and then, using implicit functions to add details to the framework. The result was that with a little bit training data, it could achieve excellent generalization effects, solving the problem of previous methods' too much reliance on large-scale data [6]. Around the same time, Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) brought new tools to from the broader field of 3D reconstruction. They not only reconstructed shapes, but could also model colors and lighting, and they can produce realistic images from any perspectives [7]. 3DGS is quite efficient, can achieve real-time rendering [8]. By 2025, Large Animatable Human Reconstruction Model (LHM) walked out a new path, using large models for feed-forward inference. There is no need to optimize for individual samples. With just one input image, a 3D animated character model can be output in just a few seconds [9]. Compared to methods from previous years, the efficiency has once reached a new level.

So far, most existing reviews probably tend to focus on introducing cutting-edge technologies, lacking causal analysis on "why technologies evolve", and rarely being able to connect the entire evolutionary logic from parametric templates to implicit functions and then to neural rendering. Therefore, this article attempts to link together the main methods of single-image 3D human reconstruction in the past decade, and answer three main questions, what each method has solved, what limitations it has left, and what the driving force behind the technological changes is.

The review is structured as follows. First, it introduces the representation methods for three-dimensional human bodies (parametric models, voxels, implicit functions, neural fields). Second, it organizes the evolution process according to the timeline. Next, it compares the experimental data of the three papers: PIFu, PIFuHD, and Geo-PIFu. Last but not least, it discusses the current challenges and future directions.

2. Different ways of storing 3D human models in computers

2.1. Parametric model method

SMPL is a model for statistical human shapes and pose, and does not directly perform "reconstruction". It uses shape parameters β to control height, body fatness, and posture parameters θ to control joint angles, transforming a standard human template into different body types and postures [1].

At the experimental level, Human Mesh Recovery (HMR) is a representative work that applies SMPL for single-image reconstruction [10]. HMR uses an end-to-end neural network to directly regress the shape parameters β and pose parameters θ of SMPL from the RGB image, and then fits them onto the SMPL template to output the complete human mesh. This method has been evaluated on standard datasets such as Human3.6M and MPI-INF-3DHP. The experiments show that the end-to-end regression framework has a significant improvement in pose estimation accuracy compared to traditional optimization methods (such as SMPLify). However, the output of HMR is a nude grid and cannot restore the details of the clothes. This indicates that the parametric model can ensure "being human", but cannot achieve "being this person".

2.2. Voxel method

DeepHuman is a representative work of voxel representation [2]. It divides the three-dimensional space into uniform voxel grids, allowing the neural network to directly predict whether there is a human in each voxel from a single RGB image. To reduce the uncertainty in the reconstruction of invisible areas, the method additionally adds the dense semantic representation generated by the SMPL model as auxiliary information.

At the experimental level, DeepHuman trained and tested based on its own built THuman Dataset (consisting of approximately 7,000 real human models). Although the paper did not report standardized quantitative indicators such as the Chamfer Distance on public benchmarks, the qualitative results and ablation experiments indicated that the voxel-based method could restore more geometric details than the pure parametric method. The limitations were also obvious, the voxel grid resolution was limited to around 256^3 . If making it more detailed, the computational load will be too much to handle. As a result, the fine structures such as clothing folds and fingers could not be restored. This directly led to the emergence of implicit function methods.

2.3. Implicit function method

PIFu is the pioneer of implicit functions [3]. Instead of directly predicting voxels, it trains a neural network to determine "whether any point in space is on the human body surface", and assists judgment through pixel alignment, extracting corresponding 2D image features for each 3D point.

Experimentally, PIFu reported a Chamfer Distance of 3.573 cm, P2S of 0.186 cm, and a normal error of 1.483 on the THuman2.0 dataset. This number is not impressive, but it is the first method that reconstructs dressed human bodies to conduct a quantitative evaluation on an open benchmark. PIFuHD improved the input image resolution from 512 to 1024 on the basis of PIFu, deepened the network, and reduced the Chamfer to 3.237 cm on the CAPE dataset, with clearer finger and facial details [4]. Geo-PIFu added global geometric constraints to the network, and compared to PIFu on public datasets, decreased the Chamfer and P2S by 42.7% and the normal error by 19.4%, being more stable in complex postures [5]. The evolution logic is clear: PIFu addresses "can it make

clothes", PIFuHD addresses "can it be clearer", and Geo-PIFu addresses "can it be more stable". It should be noted that the datasets used by the three are different, so the values cannot be directly compared horizontally.

2.4. Neural fields and real-time rendering method

NeRF uses neural networks to map the three-dimensional space into color and density values, enabling the rendering of realistic new images from any perspectives [6]. Unlike previous methods that only reconstructed "shapes", NeRF can model and shape appearance at the same time. 3DGS explicitly represents the scene using Gaussian ellipsoids, achieving real-time rendering (361 FPS) while maintaining state-of-the-art visual quality [7].

At the human body reconstruction level, recent works mainly use 3DGS as the representation method, training large models for feed-forward inference. A single image input can output related 3D characters in seconds. The advantage of 3DGS in real-time interaction is driving this field from "offline rendering" to "real-time availability". However, it should be noted that both NeRF and 3DGS as representation methods require multiple input images to optimize the scene. In single-image scenarios, additional advanced knowledge is still needed to make up for the information loss.

3. The evolution of technology, from rigid models to deep learning

From 2015 to 2018, during the early stage of template-based modeling, the emergence of SMPL introduced the natural idea of estimating parameters from images [1]. HMR used neural networks to directly regress the parameters themselves. SMPLify used optimization methods for fitting. Although the final results were stable, they were unable to reconstruct the details of the clothes, and the limitations were obvious [8].

In 2019-2020, DeepHuman was the first to attempt end-to-end deep learning, enabling the network to transform images directly into predictions of objects. However, its disadvantage was limited resolution [2]. PIFu instead used an implicit function combined with pixel alignment method, which not only achieved the first reconstruction of the body wearing clothes with wrinkles, but also broke out the resolution constraint [3]. PIFuHD successfully increased the resolution from 512 to 1024 using a double-layer structural network, and the details became more clearer [4]. Geo-PIFu is responsible for adding global geometric constraints. Now, it is more stable in representing complex action behaviors such as bending over and arms crossing [5].

From 2022 to 2026, ICON integrated SMPL and implicit functions, successfully achieving wonderful generalization results with only 12% of the training data [9]. LHM performed feed-forward inference through multi-modal Transformer, inputting a single image and outputting an animated 3D character within seconds [10]. PRADA3D also reconstructed 3D Gaussian virtual humans from single videos, achieving real-time rendering at approximately 80 frames per second. This marked the field's transition from offline single-machine fine reconstruction to online real-time usability.

4. Experimental data analysis

During the experimental process, the researchers identified three universal indicators to measure the accuracy of 3D reconstruction.

Chamfer Distance: Measures the two-way average distance between two point cloud sets, that is, the distance between points in the reconstructed model and the real model. The lower the value, the

higher the geometric similarity between the reconstructed mesh and the real mesh.

P2S (Point-to-Surface Distance): The one-way average distance from a point on the real grid to the surface of the predicted grid. That is the vertical distance from the reconstructed model to the real model. The lower the value, the higher the reconstruction accuracy.

Normal Error: Measures the difference between the normal direction of the reconstructed surface and the normal direction of the actual surface. The lower the value, the better the surface details are restored.

Table 1. Quantitative evaluation metrics for single-image 3D human body reconstruction methods

Method	Datasets	Chamfer(cm)↓	P2S(cm)↓	Normal Error↓
PIFu [3]	THuman2.0	3.573	0.186	1.483
PIFuHD [4]	CAPE	3.237	-----	-----
Geo-PIFu [5]	Public datasets	↓42.7%*	↓42.7%*	↓19.4%*

Table Note: Values marked with "—" are those not reported in the original paper. Values marked with " * " indicate the reduction compared to PIFu. The datasets used for PIFu and PIFuHD differ; therefore, the values cannot be directly compared.

As shown in Table 1, PIFu achieves a Chamfer distance of 3.573cm, P2S of 0.186cm, and a normal error of 1.483 on THuman2.0. While the numbers are not outstanding, it is the first method to quantitatively evaluate the reconstruction of the dressed human body on an open benchmark. Its significance lies in pioneering the implicit function approach [3]. PIFuHD reduces the Chamfer distance to 3.237cm on CAPE; although the reduction is not large, the qualitative difference is significant. When increase the input resolution from 512 to 1024, fingers changed from a "mass" to distinguishable joints [4]. Compared to PIFu, Geo-PIFu reduces Chamfer and P2S by 42.7% and the normal error by 19%.4%, mainly due to the introduction of global geometric features, which is more stable under complex poses [5].

The evolution logic of PIFu, PIFuHD, and Geo-PIFu is clear. PIFu addressed whether clothes could be made, PIFuHD addressed whether they could be clearer, and Geo-PIFu addressed whether they could be more stable. Since the three use different datasets, their numerical values cannot be directly compared horizontally. What is truly valuable is that how much each method has improved compared to its own baseline.

Because Deephuman did not report standardized metrics such as Chamfer/P2S on public benchmarks, it was not included in the quantitative comparison. But its main contribution is the methodological, which is the first to demonstrate that end-to-end deep learning can be used for single-image 3D human reconstruction [2].

Quantitative metrics (such as Chamfer Distance and P2S) do reflect reconstruction accuracy, but they are not all-encompassing; realistic visual dynamics are also important. Sometimes two methods have similar method values, yet the visual effects differ greatly to the eyes. For example, one reconstructs a face naturally, while another has similar numerical values but looks uncomfortable. Therefore, some good papers generally report both qualitative and quantitative results. Quantitative results provide "scientific evidence," while qualitative results provide "visual judgment."

5. Current challenges and future prospects

5.1. Current challenges

Nowadays, although 3D human body model reconstruction technology has become more mature, this field still faces three core challenges. First, data scarcity. The cost of collecting 3D human body data is significantly higher than that of 2D images. The existing datasets are limited in scale, which limits the generalization capability of data-driven approaches. Second, inconsistent evaluation standards. Different methods adopt different datasets and evaluation protocols, making it difficult to compare quantitatively. Third, the contradiction between accuracy and speed. PIFuHD offers high quality but is computationally slow, while LHM is fast but relies on large-scale training data. This conflict is particularly prominent in real-time application. However, there are two fundamental problems. One is the reconstructed human "cannot move", because most current methods only produce static models and lack the ability to model motion sequences. The other is the clothing "does not move". Even if a human wearing clothes is reconstructed, the clothes are "fixed" on the body and cannot produce natural folds or sliding effects with movement.

5.2. Future direction: deep rendering of action sequences and clothing dynamic details

Modeling and rigging for action series. Most current single-image reconstruction methods output either a static mesh or a neural field, which can pose but lack the ability to model continuous motion sequences. To let the reconstructed virtual human "move," researchers must address three key challenges. They are consistency in pose driving, preservation of temporal continuity, and physical plausibility. The video-based reconstruction methods possess advantages in this regard. However, it remains an open challenge that infer the character's behavior across a motion sequence from a single static image.

Deep rendering of clothing dynamic details. This is an issue that is one of the most overlooked yet most critical issues affecting realism. In the real world, clothing undergoes complex deformations—such as stretching, wrinkling, sliding, and drape—when moving. Existing methods either ignore clothing dynamics or employ simple deformation models, resulting in stiff visual effects. To make the reconstructed clothed human truly lifelike, it is essential to learn deformation priors for different materials (e. g., cotton, silk, leather). So far, there is currently a lack of large-scale clothing dynamic datasets; combining synthetic data with real data may be a breakthrough direction. Additionally, it's necessary to integrate physical simulation with neural rendering, while traditional fabric simulation can generate highly realistic clothing dynamics. It suffers from high computational costs, and the neural rendering methods (such as NeRF and 3DGS) can efficiently generate realistic images but often lack physical law modeling. Thus, combining the two worthwhile to explore. Finally, there is no doubt that modeling the interaction between clothing and the body is necessary. The clothing and the body surface have friction, sliding and collision, how to maintain real-time rendering efficiency simultaneously is one of the ultimate challenges of depth rendering.

6. Conclusion

Over the past decade, three-dimensional human body reconstruction from single images has undergone rapid development, transitioning from parameterized fixed templates to voxels, and from implicit functions to neural rendering. The implicit function-based approaches represented by the PIFu series have laid the technical foundation—using neural networks to implicitly represent 3D

shapes, clearly bypassed the voxel resolution bottleneck. According to experimental data, the resolution contribution of PIFuHD and the global geometric constraints of Geo-PIFu have continuously improved the accuracy and stability of the reconstruction. As technologies in this field have begun to be officially extended to the broader public, the feed-forward reasoning and real-time rendering capabilities of large language models have progressively enhanced their efficiency and practical applicability. Although the problems of data scarcity and weak generalization ability still exist, with the gradual application of multi-model and diffusion models, researchers can confidently believe that the problems will eventually be solved. Continuously improving to build a complete technical system.

References

- [1] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (pp. 851-866).
- [2] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.
- [3] Zheng, Z., Yu, T., Wei, Y., Dai, Q., & Liu, Y. (2019, October). Deephuman: 3d human reconstruction from a single image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7738-7748). IEEE.
- [4] Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., & Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2304-2314).
- [5] Saito, S., Simon, T., Saragih, J., & Joo, H. (2020, June). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 81-90). IEEE.
- [6] Xiu, Y., Yang, J., Tzionas, D., & Black, M. J. (2022, June). Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13286-13296). IEEE.
- [7] He, T., Collomosse, J., Jin, H., & Soatto, S. (2020). Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33, 9276-9287.
- [8] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99-106.
- [9] Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., ... & Bo, L. (2025, October). Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 14184-14194). IEEE.
- [10] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *arXiv*. arXiv preprint arXiv:2308.04079.