Human Body Reconstruction Based on Multiple Data Sources

Wanfa Zhang 36920231153263 AI_class,[1] Yaming Yang 23020231154245 CS_class,[2] Xinyuan Du 24520230257501 CS_class,[2] Fan Lei 36920231153206 AI_class,[1] Jianhong Lan 36920231153205 AI_class[1]

[1]Institute of Artificial Intelligence
[2]School of Informatics
Xiamen University
Xiamen, China

Abstract

Human body reconstruction primarily relies on single RGB images as the data source, but their 2D nature limits depth accuracy. Point cloud-based reconstruction provides depth and normals but often results in sparse, dissimilar meshes. We propose a novel approach using multi-modal data (images and point clouds) for human reconstruction.Our method involves three steps: 1) feature extraction from images and point clouds to generate a parameterized human model (SMPL), 2) designing a loss function to predict precise human body normals from images and the SMPL, and 3) using IF-NET++ to bridge normal prediction gaps and achieve a complete mesh reconstruction.Our multi-modal human reconstruction overcomes single-modal limitations, supported by quantitative and qualitative experiments demonstrating its effectiveness.

Introduction

Highly realistic virtual humans are poised to play a pivotal role in augmented and mixed reality, forming a crucial foundation for the concept of the ”metaverse.” This advancement promises to support remote presentations, collaborations, education, and entertainment. To achieve this, new tools are needed to effortlessly create and animate 3D virtual characters. Traditionally, this has required substantial artistic effort and expensive scanning equipment.

Current approaches to human body reconstruction fall into three primary categories: explicit reconstruction based on parameter models like SMPL (Loper et al. 2023), implicit reconstruction based on models like distance fields and occupancy fields, and more recent NeRF-based methodsp (Mildenhall et al. 2021; **?**) for human body reconstruction. However, most of these methods rely on single-modal data sources, such as images, stereo images, or RGBD images, resulting in sparse and imprecise information. Recent advancements in LiDAR technology have led to the acquisition of increasingly dense point cloud data, offering more accurate depth and normal information. This has opened the door to the fusion of point cloud data with image data for improved human body reconstruction, yielding more accurate results.

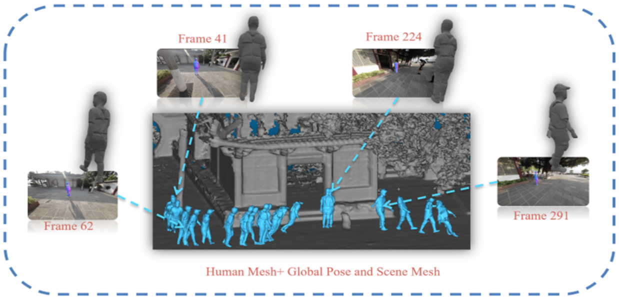

Figure 1: Visualizations were conducted on the SLPOER4D dataset, where the generated Meshes were augmented with global trajectories and projected back into the scene maps.

While methods like PIFu(HD) (Saito et al. 2019, 2020) can reconstruct 3D human figures in attire with unconstrained topological structure, they often overfit to the poses seen in training data and lack explicit knowledge of human body structure, leading to unrealistic limb shapes. Explicit body models can be used to regularize implicit models, but this introduces topological constraints, limiting their ability to generalize to novel clothing styles and compromising shape detail.

To address these challenges, we propose a multi-modal human body reconstruction method that combines image and point cloud data. Leveraging state-of-the-art frameworks, we integrate our point cloud data into image stream processing. Specifically, we utilize the accurate depth and normal information from the point cloud to provide more precise SMPL parameter estimation, overcoming depth uncertainty in SMPL parameter estimation from images. Furthermore, our point cloud data includes information about both the human body and clothing, enabling our implicit model to provide more comprehensive depth information for both the human body and its attire, beyond the prior information from SMPL parameters.

Our method comprises three primary steps: First, we extract features from images and point clouds to generate a parameterized human model (SMPL). Second, we design a loss function to accurately predict human normals from images and SMPL parameters. Finally, we employ IF-NET++ to bridge the gap between predicted normals and achieve a complete mesh reconstruction. Our multi-modal human body reconstruction overcomes the limitations of single-modal data, as demonstrated through both quantitative and qualitative experiments.

Related work

Clothed Human Reconstruction from Images. In the domain of computer vision and computer graphics, image-based clothed human reconstruction has garnered significant attention due to its applications in areas like virtual try-on, character animation, and digital fashion. This research area aims to create 3D models of clothed humans directly from 2D images or image sequences, allowing for realistic and dynamic representation of human subjects in various outfits. The following sections will delve into three key categories of related work in image-based clothed human reconstruction, providing an overview of the different methodologies and approaches that have been developed to address this challenging problem.

Approaches Based on Explicit 3D Shape Modeling. In the field of 3D human reconstruction, diverse methodologies have been employed. Explicit-shape-based techniques utilize either a mesh-based parametric body model (Joo, Simon, and Sheikh 2018; Loper et al. 2023; Romero, Tzionas, and Black 2022) or nonparametric representations like depth maps (Gabeur et al. 2019) or point clouds (Zakharkin et al. 2021) for the creation of 3D human models. Many approaches focus on the estimation or regression of 3D body meshes from RGB images while neglecting clothing details (Feng et al. 2021; Kocabas et al. 2021; Sun et al. 2022; Zhang et al. 2023, 2021). To account for the complexities of clothed human forms, another line of research introduces 3D offsets on top of the body mesh (Pons-Moll et al. 2017). This approach seamlessly integrates with existing animation pipelines, as it inherits the hierarchical skeleton and skinning weights from underlying statistical body models. However, the ”body+offset” approach falls short in adequately modeling loose clothing, such as dresses and skirts, which significantly deviate from the body’s topology. In an effort to enhance topological flexibility, some methods reconstruct 3D clothed humans by recognizing the type of clothing and employing the appropriate models for reconstruction (Jiang et al. 2020). However, scaling up this ”cloth-aware” approach to accommodate a wide range of clothing styles is a nontrivial task, limiting its applicability to handling diverse outfit variations encountered in real-world scenarios.

Approaches Based on Variations in Implicit-function. Implicit-function-based techniques provide a versatile approach for representing diverse 3D clothed human shapes. Methods like SMPLicit (Corona et al. 2021), and DIG (Li et al. 2022) employ neural distance fields to create generative clothing models from 3D clothing datasets. When given an image, these methods reconstruct clothed humans by estimating a parametric body and optimizing the latent space of the clothing model. However, results often suffer from misalignment with the image and lack of geometric detail. PIFu (Saito et al. 2019) introduces pixel-aligned implicit human shape reconstruction, while PIFuHD (Saito et al. 2020) enhances geometric detail with a multi-level architecture and normal maps from RGB images. These methods do not utilize human body structure knowledge, leading to overfitting to specific body poses in training data, limiting their generalization to new poses. To address these issues, some methods introduce geometric priors for regularization. GeoPIFu (He et al. 2020) models a coarse shape of volumetric humans, and PINA (Dong et al. 2022), and S3 (Yang et al. 2021) use depth or LIDAR information to improve shape regularity. Another approach combines parametric body models with implicit representations for the best of both worlds. PaMIR (Zheng et al. 2021), ARCH (Huang et al. 2020), ARCH++ (Huang et al. 2020) use SMPL or 3DMM to enhance their reconstructions.

Approaches Based on NeRF for Human Reconstruction. NeRF (Neural Radiance Fields) (Mildenhall et al. 2021) has been a source of inspiration for research in 3D human reconstruction. Human NeRF techniques have emerged to generate high-quality views and poses of 3D humans using multi-view or monocular human videos. For instance, Neural Body (**?**) applies sparse convolutions to model radiance volumes. Meanwhile, other approaches model human NeRF in canonical spaces (**?**Su et al. 2021) using SMPL (Skinned Multi-Person Linear) body model weights or optimizing these weights with appearance information. While these methods yield impressive results, they often require extensive computation and dense observations, making them less efficient. To overcome this challenge, there is a growing interest in developing generalizable human NeRF techniques (Zhao et al. 2022; Kwon et al. 2021). These methods reduce the need for extensive observations and achieve reconstruction with a single forward pass. This work aims to contribute to the advancement of generalizable human NeRF, focusing on a more complex task: recovering animatable human NeRF from a single image.

Method

Normal Reconstruction by RGB and Point cloud

Point Cloud Feature Embedding. For point cloud feature extraction, we employ the well-established PointNet++ architecture. Diverging from conventional approaches that involve voxelization of point clouds followed by the application of standard neural networks, PointNet++ directly processes raw 3D point clouds. This characteristic makes it particularly advantageous for preserving the inherent spatial information of the point cloud.

PointNet++ operates without the need for voxelization, accepting raw point cloud data as input, where each point is represented by its spatial coordinates and additional features. The network employs a hierarchical architecture with multiple set abstraction (SA) and feature propagation (FP) layers. The SA layers hierarchically extract features at different scales, allowing the model to capture both fine and coarse details in the point cloud. The sampling mechanism is employed to downsample points, and a grouping operation aggregates features within local regions. This enables the model to efficiently capture local structures. Through iterative feature aggregation, the network synthesizes global and local information, ensuring that the resulting feature representation is comprehensive and informative.

The utilization of PointNet++ proves instrumental in extracting rich and discriminative features from the raw point cloud. This feature richness is crucial for generating high-quality normal maps in subsequent stages of our reconstruction pipeline.

ResNet-based Fusion for Normal Map Generation. In our reconstruction pipeline, we incorporate ResNet-based fusion to synergize RGB image and point cloud data. Initially, RGB images undergo down-sampling to extract high-dimensional features, which are fused with PointNet++-extracted point cloud features. These fused features are processed through ResNet, known for its residual learning framework that effectively handles deep networks.

ResNet captures intricate patterns and hierarchical features, making it well-suited for our task. Post-ResNet processing involves up-sampling and activation to generate a higher resolution representation. The final normal map is derived from these processed features, detailing the surface orientations of the reconstructed 3D model.

This ResNet-based fusion strategy plays a pivotal role in integrating multi-modal information, enhancing normal map quality, and benefiting from ResNet's advantages in handling complex patterns within our reconstruction approach, see Figure 2.

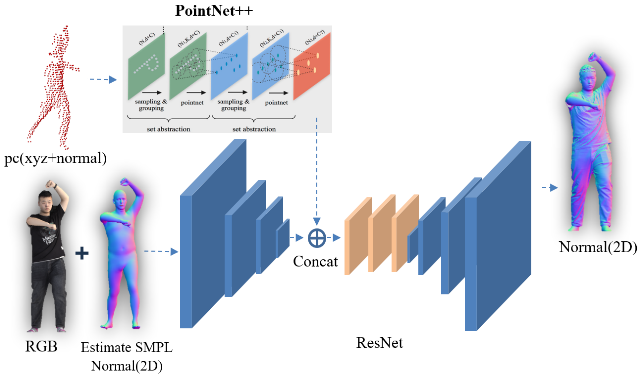

Figure 2: The diagram illustrates our network architecture for normal map generation. The inputs include point cloud data, RGB images, and the predicted SMPL model. These modalities undergo a feature fusion process, combining point cloud features extracted by PointNet++ with down-sampled features from RGB images. The fused features are then processed through a ResNet network, resulting in the generation of the corresponding normal map.

Rough mesh reconstruction of human body

In our research, we adopted a methodology similar to (Xiu et al. 2023) for rough mesh reconstruction of human body. Three key aspects ensured: first, high-frequency surface details are aligned with the predicted clothed normal maps; second, low-frequency surface variances are consistent with those from the SMPL-X model; and third, the depth profiles of the front and back silhouettes are closely matched.

This approach use neural networks to deduce implicit surfaces from normal maps, explicitly modeled the depth-normal relationship through variational normal integration methods (Cao et al. 2022; Quéau, Durou, and Aujol 2018). The recent bilateral normal integration (BiNI) method (Cao et al. 2022) was customized for full-body mesh reconstruction, incorporating a coarse prior, depth maps, and silhouette consistency.

A depth-aware silhouette-consistent bilateral normal integration (d-BiNI) method (Xiu et al. 2023) to fulfill these conditions. This approach jointly optimizes the front and back clothed depth maps, significantly improving the accuracy of the depth mapping.

Objective function includes several components:

$$\begin{aligned}\min_{\hat{Z}_F^c, \hat{Z}_B^c} \; & L_n(\hat{Z}_F^c; \hat{N}_F^c) + L_n(\hat{Z}_B^c; \hat{N}_B^c) + \\ & \lambda_d L_d(\hat{Z}_F^c; \hat{Z}_F^b) + \lambda_d L_d(\hat{Z}_B^c; \hat{Z}_B^b) + \\ & \lambda_s L_s(\hat{Z}_F^c, \hat{Z}_B^c),\end{aligned} \quad (1)$$

the BiNI loss term L_n, a depth prior L_d for front and back depth surfaces, and a silhouette consistency term L_s. The depth prior

$$L_d(\hat{Z}_i^c; \hat{Z}_i^b) = |\hat{Z}_i^c - \hat{Z}_i^b|_{\Omega_n \cap \Omega_z}, \quad i \in \{F, B\}, \quad (2)$$

derived from the SMPL-X body mesh, helps align the front and back surfaces coherently, addressing the challenge of unifying these surfaces into a complete body structure. The silhouette consistency term:

$$L_s(\hat{Z}_F^c, \hat{Z}_B^c) = |\hat{Z}_F^c - \hat{Z}_B^c|_{\partial\Omega_n}, \quad (3)$$

is particularly crucial in maintaining the physical integrity of the reconstructed depth maps, preventing undesirable artifacts and improving overall reconstruction quality.

By integrating these components, this methodology makes significant technical advancements beyond the basic BiNI approach (Cao et al. 2022), particularly in terms of depth accuracy and surface coherence.

Poisson mesh reconstruction

We drew upon the methodology described in (Xiu et al. 2023), particularly when dealing with complex poses that result in self-occlusions. For simple poses, the straightforward fusion of d-BiNI surfaces suffices, as previously demonstrated by FACSIMILE (Smith et al. 2019) and Moduling Humans(Gabeur et al. 2019). However, this approach is insufficient for poses with self-occlusion, which leave large portions of the surface incomplete and prone to the creation of blobby artifacts when applying Poisson Surface Reconstruction (PSR)(Kazhdan, Bolitho, and Hoppe 2006).

To address this, we employed a two-pronged strategy, first introducing an approach we termed ECON_{EX}. This technique builds upon PSR (Kazhdan, Bolitho, and Hoppe 2006) by incorporating the estimated SMPL-X body model to infill missing surfaces. While ECON_{EX} effectively avoids missing limbs, it does not fully resolve issues of surface coherence for clothing and hair, due to the inherent differences between the SMPL-X model and actual garments or hair.

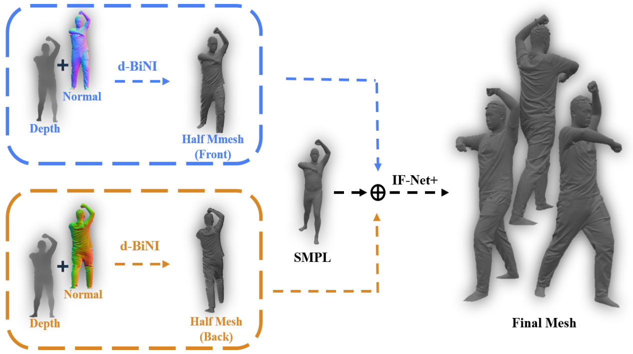

Figure 3: The figure demonstrates the optimization process of depth and normal maps through the d-BiNI method, elevating them into 3D space (depicted as half of the mesh). Subsequently, the front and back meshes, along with the predicted SMPL input, are fed into IF-Net+ to obtain the final mesh result.

We further refined our reconstruction by implementing inpainting with an improved implicit-function model, IF-Nets+ (Chibane, Alldieck, and Pons-Moll 2020), which enhances the generic shape completion capabilities of IF-Nets(Chibane, Alldieck, and Pons-Moll 2020). IF-Nets+ is adapted to the SMPL-X model to accommodate variations in pose, trained on voxelized depth maps and body meshes, and supervised with ground-truth 3D shapes. This training process includes random masking of depth maps to increase the model's robustness to occlusions. For inference, we input the estimated depth maps and body mesh into IF-Nets+ to produce an occupancy field, which we then convert into the inpainted mesh $\mathcal{R}_{IF}$ via the Marching cubes(We 1987) .

Our final mesh, denoted as $\mathrm{ECON}_{\mathrm{IF}}$, results from the PSR-based integration of d-BiNI surfaces, side-view and occluded regions from $\mathcal{R}_{IF}$, and, when necessary, elements from the SMPL-X body mesh to rectify poorly reconstructed areas like hands or face. Although $\mathcal{R}_{IF}$ alone yields a complete human mesh, it tends to smooth out finer details. Hence, in our process, only the side-views and occluded parts of $\mathcal{R}_{IF}$ are fused in the Poisson step to better preserve the detail captured by d-BiNI. This nuanced approach, which aligns with the strategy delineated in (Xiu et al. 2023), ensures the preservation of high-fidelity details in our final reconstructed models, as evidenced by our evaluation metrics. Our pipiline is listed in Figure 3.

Experiments

Datasets

Typically, to obtain three-dimensional point cloud data of the human body, it is necessary to use a laser radar to scan the target body from multiple angles. The point cloud results obtained from scans at different angles are then stitched together to generate the final three-dimensional point cloud data. The dataset used in this experiment comes from THuman2.0, where the three-dimensional human body data is stored in mesh files. As we cannot directly convert mesh-format files into point cloud format, we need to simulate the laser radar scanning process on the mesh data from THuman2.0 to obtain the desired point cloud files. For this simulation, we selected the OS1 high-definition imaging radar from OUSTER. In terms of implementation, we drew inspiration from the principles of ray tracing. Using the human body mesh data as the origin, we simulated the rotation of the laser radar around the horizontal plane of the human body mesh data to obtain point cloud data at different angles.

Training on Thuman2.0. We conducted tests on the Thuman2.0 dataset, specifically selecting 525 high-quality three-dimensional human texture scans. Each high-resolution scan was divided into 36 viewpoints at 10-degree intervals. Additionally, corresponding SMPL-X and point cloud data were generated based on images. Both PIFUHD and PaMIR were retrained on this training set.

Evaluation on CAPE. We conducted tests on the CAPE human dataset. Specifically, we employed a similar processing method as the training set, capturing viewpoints every 120 degrees. Corresponding SMPL-X and point cloud data were generated and utilized in the inference process.

Metrics

Chamfer and P2S distance. We evaluate the commonly used Chamfer and P2S distance between ground-truth and reconstructed meshes to capture large geometric errors, for instance, occluded parts or wrongly positioned limbs.

L2. For a more detailed evaluation of the reconstructed meshes, particularly concerning the precision of local surface details and the consistency of projections from the input image, we report the L2 norm of the normal differences. This involves rendering normal images from both the reconstructed and ground-truth surfaces and calculating the L2 error. We perform this assessment by rotating a virtual camera around the meshes at intervals of 0°, 90°, 180°, 270° relative to a frontal view.

Evaluation

Quantitative evaluation. We compared our method with body-agnostic approaches such as PIFuHD (Saito et al. 2020) and body-aware methods like PaMIR (Zheng et al. 2021). To ensure a fair comparison, we re-implemented PIFuHD and PaMIR. As shown in Table 1, the optimization method based on SMPL-X achieved the best performance across the three metrics. The regression-based IF-Net+ method outperforms PaMIR and PIFuHD in terms of chamfer and P2S metrics. Additionally, it surpasses PIFuHD in the Normal metric but is inferior to PaMIR. See Figure 4.

Qualitative evaluation. For a more intuitive comparison with PIFuHD and PaMIR, we selected challenging datasets, including instances with challenging poses and loose clothing. Notably, for datasets with loose clothing, both PIFuHD and PaMIR exhibit limitations in accurately capturing high-frequency details. Additionally, in challenging pose scenarios, PIFuHD experiences instances of failure.

Additionally, to validate the effectiveness of our method on in-the-wild datasets, we conducted tests on the

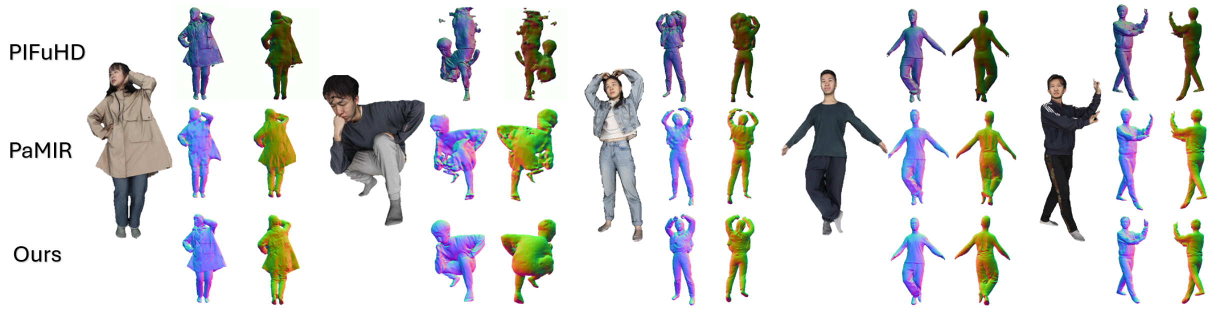

Figure 4: We visualized several result sets, including scenarios with loose clothing, challenging poses, and normal poses. From top to bottom, the order is PIFuHD, PaMIR, and Our Method. The images display the frontal mesh results on the left and the dorsal mesh results on the right.

Table 1: *Method indicates the re-implemented approach, EX subscript signifies the utilization of optimization methods, IF denotes the application of regression methods, and ”OOD” represents “out-of-distribution.

Methods	OOD poses(CAPE)		
	Chamfer↓	P2S↓	Normals↓
PaMIR*	1.023	1.133	0.0422
PIFuHD	3.767	3.591	0.0994
$Ours_{IF}$	1.134	1.122	0.0457
$Ours_{EX}$	**1.066**	**1.083**	**0.0413**

SLPOER4D dataset. Meshes were generated based on images and point clouds, and leveraging the provided global body trajectories in the dataset, the Meshes were registered to the scenes, resulting in the effects shown in Figure 1.

Ablation study. In the ablation study conducted in the second part of our pipeline, we introduced point clouds to supplement depth information in the presence of the original SMPL human depth priors. A qualitative analysis reveals that, compared to scenarios where only human depth priors are utilized, the reconstruction of clothing in conjunction with the human body is significantly more accurate. Quantitative analysis results are provided in Table 2.

Table 2: The first row corresponds to scenarios without the inclusion of point cloud depth information, while the second row represents scenarios incorporating point cloud depth information.

Methods	OOD pose(CAPE)		
	Chamfer↓	P2S↓	Normals↓
w/o pc_depth	1.123	1.258	0.0622
w/ pc_depth	**1.066**	**1.083**	**0.0413**

Conclusion

In summary, we have proposed a multi-modal data fusion framework for human mesh reconstruction. Leveraging the characteristics of point cloud data and the semantic richness of image data, our framework adeptly restores comprehensive human body surfaces. Relative to single-modal methods, we harness the distinctive features of point cloud data, establishing a robust coupling relationship between the reconstruction of human and clothing surfaces. This coupling facilitates a well-resolved depth ambiguity, addressing the challenges of multiple interpretations inherent in the depth data. Notably, our framework has been successfully tested in in-the-wild scenarios, showcasing commendable reconstruction results. This capability holds promise for providing richer annotation data in large-scale multi-modal datasets.

References

Cao, X.; Santo, H.; Shi, B.; Okura, F.; and Matsushita, Y. 2022. Bilateral normal integration. In *European Conference on Computer Vision*, 552–567. Springer.

Chibane, J.; Alldieck, T.; and Pons-Moll, G. 2020. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6970–6981.

Corona, E.; Pumarola, A.; Alenya, G.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11875–11885.

Dong, Z.; Guo, C.; Song, J.; Chen, X.; Geiger, A.; and Hilliges, O. 2022. PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20470–20480.

Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, 792–804. IEEE.

Gabeur, V.; Franco, J.-S.; Martin, X.; Schmid, C.; and Rogez, G. 2019. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2232–2241.

He, T.; Collomosse, J.; Jin, H.; and Soatto, S. 2020. Geopifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33: 9276–9287.

Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3093–3102.

Jiang, B.; Zhang, J.; Hong, Y.; Luo, J.; Liu, L.; and Bao, H. 2020. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 18–35. Springer.

Joo, H.; Simon, T.; and Sheikh, Y. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8320–8329.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 0.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11127–11137.

Kwon, Y.; Kim, D.; Ceylan, D.; and Fuchs, H. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34: 24741–24752.

Li, R.; Guillard, B.; Remelli, E.; and Fua, P. 2022. Dig: Draping implicit garment over the human body. In *Proceedings of the Asian Conference on Computer Vision*, 2780–2795.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Pons-Moll, G.; Pujades, S.; Hu, S.; and Black, M. J. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4): 1–15.

Quéau, Y.; Durou, J.-D.; and Aujol, J.-F. 2018. Normal integration: a survey. *Journal of Mathematical Imaging and Vision*, 60: 576–593.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and 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*, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 84–93.

Smith, D.; Loper, M.; Hu, X.; Mavroidis, P.; and Romero, J. 2019. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5330–5339.

Su, S.-Y.; Yu, F.; Zollhöfer, M.; and Rhodin, H. 2021. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34: 12278–12291.

Sun, Y.; Liu, W.; Bao, Q.; Fu, Y.; Mei, T.; and Black, M. J. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13243–13252.

We, L. 1987. Marching cubes: A high resolution 3d surface construction algorithm. *Comput Graph*, 21: 163–169.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 512–523.

Yang, Z.; Wang, S.; Manivasagam, S.; Huang, Z.; Ma, W.-C.; Yan, X.; Yumer, E.; and Urtasun, R. 2021. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13284–13293.

Zakharkin, I.; Mazur, K.; Grigorev, A.; and Lempitsky, V. 2021. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14718–14727.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11446–11456.

Zhao, F.; Yang, W.; Zhang, J.; Lin, P.; Zhang, Y.; Yu, J.; and Xu, L. 2022. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7743–7753.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184.