3D Human Digitization with High-fidelity Textures from Single Image

Xin Wang*
School of Software, Beihang University
22373114@buaa.edu.cn

Nanxiang Jiang*
Shen Yuan Honors College, Beihang University
jiangnx@buaa.edu.cn

Figure 1. 3D Reconstruction results for single smartphone captured photos. Our approach emphasizes pose completeness and fine-grained accuracy, with particular attention to the coherent propagation of front and back full-body textures, resulting in highly fidelity outcomes.

Abstract

Reconstructing high-quality 3D models of clothed humans from single images for is of crucial importance for real world applications. Despite recent advancement in accurately reconstructing humans in complex poses or with loose clothing from in-the-wild images, the problem of predicting textures for unseen areas and generating full-body high-fidelity texture maps remains a significant challenge. A key limitation of previous methods is their incompatible meshs and insufficient prior guidance in transitioning from 2D to 3D and in texture prediction. In response, we introduce a two-stage method to tackle this problem: 1) ***High-fidelity 3D human mesh contruction from single image****, which uses the low-resolution depth network to predict the global structure from a low-resolution image and uses the part-wise image-to-normal network to predict the details of the 3D human body structure. Then the high-resolution depth network merges the global 3D shape and the detailed structures to infer the high-resolution front and back side depth maps, resulting a mesh containing the global shape of a human and its details. 2)* ***Full body clothed texture prediction and refinement****, which combines a side-view decoupling transformer with a cross-attention mechanism, using SMPL-X normals as queries to effectively decouple side-view features in the process of mapping 2D features to 3D. Through extensive experiments, our approach surpasses SOTA methods in both geometry and texture reconstruction, showcasing enhanced robustness in different human bodies and complex scenarios, achieving striking results in P2S measurement. Our approach extends to practical applications such as 3D printing and human mesh building, demonstrating its broad utility in real-world applications.*

1. Introduction

Reconstructing realistic 3D human models with detailed shapes and full-body textures is a key problem in computer vision and graphics. These models are widely used in applications such as virtual and augmented reality (VR/AR), digital content creation, 3D printing, online shopping, gaming, and film production [2, 7, 9, 40]. High-quality digital humans make it possible to create immersive experi-

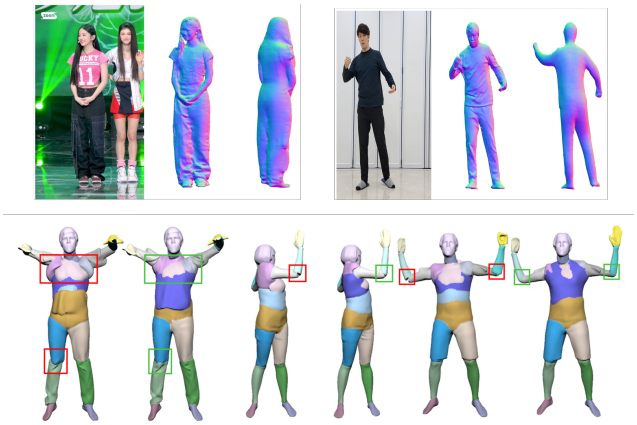

Figure 2. 2k2k [12](top) and UNIF [32](bottom) excel on reconstructing 3D human geometry, but overlook the prediction and completion of back-view textures.

ences and enable personalized content at scale. Traditional methods often rely on multi-view images and require expensive camera arrays and controlled environments to capture high-quality data. These systems are not only costly and bulky, but also require expert operation, making them impractical for everyday users like content creators and influencers [4, 21–23, 25, 26, 26, 27, 29, 33, 34, 37].

Therefore, recent researches in deep learning and computer vision has focused on reconstructing human models from a single image [2, 13, 21, 28, 31, 35, 38–41]. In recent years, a wide range of 3D reconstruction methods have emerged, notable among them are approaches based on deep implicit volumes and multiple depth maps, which represent two prominent lines of research. Among these, UNIF [32] and 2k2k [12] stand out as strong representatives, delivering impressive performance and serving as influential benchmarks in their respective categories. However, these methods primarily focus on reconstructing 3D human geometry, emphasizing depth map completeness, surface normal accuracy, and fine mesh details. While these geometric cues are crucial, they tend to overlook an essential aspect of monocular human reconstruction: the prediction and completion of back-view textures. As illustrated in Figure 2, existing methods often resort to overly simplistic solutions for the unseen back side. For instance, 2k2k directly mirrors the front-view texture onto the back, which is clearly unrealistic and unsuitable for downstream applications. Consequently, generating 3D digital humans with plausible and complete back-view textures remains an unsolved and underexplored problem.

Nevertheless, generating 3D digital humans with both front- and back-view textures from a single image holds substantial practical value, especially in everyday scenarios and commercial media applications. There is a growing demand for accessible solutions that allow users to create high-fidelity, fully textured digital avatars using only a lightweight device——such as capturing a single photo with a mobile phone. Imagine walking into a photo studio, having only your front side photographed, and receiving a hyper-realistic figurine of yourself with complete front and back textures the very next day. This kind of seamless experience would be both astonishing and transformative for personal content creation, virtual try-on, gaming, and personalized merchandising. However, current methods fall short in achieving this goal, often struggling to plausibly infer or synthesize the unseen back-view appearance.

In response to its importance and novelty, we introduce a two-stage method to tackle the problem: 1) **High-fidelity 3D human mesh contruction from single image**, which uses a low-resolution depth network to predict the global structure from a low-resolution image and uses a part-wise image-to-normal network to predict the details of the 3D human body structure. Then the high-resolution depth network merges the global 3D shape and the detailed structures to infer the high-resolution front and back side depth maps, resulting a mesh containing the global shape of a human and its details. 2) **Full body clothed texture prediction and refinement**, which combines a side-view decoupling transformer with a cross-attention mechanism, using SMPL-X normals as queries to effectively decouple side-view features in the process of mapping 2D features to 3D. Through extensive experiments, our approach surpasses SOTA methods in both geometry and texture reconstruction, showcasing enhanced robustness in different human bodies and complex scenarios, achieving striking results in P2S measurement.

To summarize, the contributions of our works are:

- **New problem.** We introduce the new research topic – generating 3D Human Digitization with High-fidelity Textures from Single Image, which demands both realistic 3D depth map and mesh generation, and full-body texture map prediction and refinement.
- **New approach.** we introduce a two-stage method to tackle the problem: High-fidelity 3D human mesh contruction from single image and Full body clothed texture prediction and refinement, which work together to reconstruct a detailed human mesh with high fidelity front textures and compatible predicted back textures.
- **New applications.** Our proposed model achieves state-of-the-art performance in both geometry and texture reconstruction, facilitating real-world applications such as 3D printing and scene building, which were challenging to achieve with previous methods.

2. Related Work

We review clothed human 3D reconstruction from single images with three basic approaches: **Parametric model-based human reconstruction**, **model-free human recon-**

struction, and other novel methods like **NeRF-based human reconstruction**.

2.1. Parametric model-based human reconstruction

Implicit representations, such as occupancy fields and signed distance functions (SDFs), offer strong topological flexibility and have demonstrated robust performance in modeling 3D clothed humans under diverse conditions, including loose garments, occlusions, and complex poses. A number of approaches leverage parametric human body priors (e.g., SMPL [24] or SMPL-X [30]) to guide 2D feature extraction and improve 3D reasoning for construction [5, 6, 8, 14, 16, 18, 32].

Among these, **SMPL** [24] (Skinned Multi-Person Linear model) is a foundational parametric model that learns to represent human body shape and pose-dependent deformations from a large corpus of registered 3D scans. It models the human body as a skinned vertex-based mesh driven by pose and shape parameters with 6890 vertices, with blend shapes that capture both identity and pose-dependent variations. A key innovation is its use of a linear function of rotation matrices to model pose-dependent deformations, allowing efficient learning and animation. SMPL not only achieves higher accuracy than earlier models like BlendSCAPE but also integrates seamlessly with standard graphics pipelines through linear or dual-quaternion blend skinning. Its generality, compatibility, and extensibility have made it a widely adopted foundation for downstream tasks such as human mesh recovery, animation, and avatar generation.

UNIF [32] introduces a unified implicit function framework for high-quality 3D human reconstruction from a single image. Instead of relying on separate surface or texture representations, it learns a shared implicit field that jointly encodes geometry and appearance within a canonical space. By leveraging pose-guided feature aggregation and a tailored volume rendering strategy, it enables the recovery of both accurate 3D shapes and realistic surface textures. This unified design simplifies the pipeline, improves consistency between geometry and texture, and achieves state-of-the-art results on in-the-wild datasets, especially under complex clothing and pose conditions.

ICON [39] builds upon parametric models like SMPL to capture the coarse body pose and shape, but further refines the surface by learning implicit functions that model clothed geometry as SDFs. By leveraging pose-aligned feature extraction and depth supervision, ICON effectively preserves high-frequency surface details such as loose clothing and wrinkles. The pipeline enables high-quality reconstruction of clothed humans while maintaining compatibility with the underlying SMPL topology, thus combining realism with structure consistency. ICON sets a new standard for geometry refinement in monocular human reconstruction, pushing the boundary of accuracy and expressiveness in real-world conditions.

2.2. Model-free human reconstruction

End-to-end pipelines [1, 3] and non-parametric forms like depth maps, normal maps, and point clouds [13, 15, 17, 19, 20, 35, 36, 38, 40] are explored for creating representations of clothed humans.

However, the end-to-end training of a volume prediction network requires a large memory footprint and computation. For this reason, these approaches apply novel methods to tackle this problem. For example, **ARCH** [19] and **ARCH++** [15] leverage the parametric SMPL model as a scaffold to guide the reconstruction but introduces a surface refinement module that predicts detailed geometry offsets in a canonical space. This design enables the model to capture complex clothing shapes and body details while ensuring that the output remains topologically compatible with animation systems. By combining image features with pose-guided priors, they produce textured 3D avatars that can be easily rigged and animated.

PIFu [35] and its variants [13, 17] learn a continuous occupancy field conditioned on per-pixel image features and 3D spatial locations, allowing it to model detailed geometry with arbitrary topology. This implicit representation enables fine-grained reconstruction of clothed human bodies, capturing subtle surface variations without relying on parametric priors like SMPL. PIFu set a new standard for image-to-geometry methods and has inspired a series of follow-up works in the area of implicit surface modeling and single-view 3D reconstruction.

2K2K [12] introduces a highly practical pipeline for reconstructing detailed 3D human models from a single high-resolution image (2048×2048), with a strong focus on efficiency, accuracy, and memory consumption. The core idea is to divide the reconstruction process into part-wise normal estimation and fusion. Specifically, the method first estimates normal maps for each body part using a part-aware encoder, guided by 2D keypoints. These part-wise normal maps are then fused into a complete normal map and used to compute a high-quality depth map. Finally, a mesh is recovered from the depth via traditional surface reconstruction algorithms. This divide-and-conquer strategy allows the model to process 2K-resolution inputs while keeping GPU memory usage low. Moreover, by focusing on part-specific features, the method preserves fine-grained geometric details, such as sharp edges and garment wrinkles, achieving high-quality results at real-time speed. The lightweight design and strong generalization ability make 2K2K well suited for deployment in practical applications such as virtual try-on and digital avatar creation.

2.3. NeRF-based Reconstruction

The rise of **Ne**ural **R**adiance **F**ields (NeRF) has seen methods [10, 11, 14] using videos or multi-view images to optimize NeRF for human form capture. These approaches can achieve impressive photorealism and geometry detail, especially in controlled settings, but often require large-scale training data, long optimization times, and multi-view consistency that limits their applicability in casual or real-time scenarios.

3. Method

Our approach comprises a two-stage pipeline for high-fidelity clothed human reconstruction from a single image. As illustrated in Fig. 3, we first reconstruct a detailed 3D human mesh using the 2K2K framework, then feed the resulting mesh into an enhanced version of SIFU for full-body texture prediction and refinement. This decoupled strategy enables both high geometric accuracy and photo-realistic texture quality, addressing common limitations in existing end-to-end single-image reconstruction systems.

3.1. High-Fidelity Mesh Construction via 2K2K

Given a single high-resolution RGB image $I \in \mathbb{R}^{2048\times2048\times3}$, we first reconstruct a clothed human mesh $M \in \mathbb{R}^{N\times3}$, where N is the number of vertices, using the 2K2K pipeline. This stage consists of three major steps:

3.1.1. Part-wise Normal Prediction

We begin by extracting body part patches guided by 2D keypoints J, which define twelve semantic parts (head, torso, limbs, feet). Each cropped region is aligned to a canonical orientation using a similarity transform M_i. This alignment normalizes pose variations and isolates body parts from background clutter.

The aligned patches are then processed by part-specific AU-Net backbones to predict double-sided (front and back) surface normal maps $\bar{n}_i$. These normal maps are inversely warped back to the image space using M_i^{-1}, and merged into a global high-resolution normal map N_h using Gaussian-weighted blending to handle boundary overlaps.

3.1.2. Low and High-Resolution Depth Estimation

To guide depth estimation, we adopt a dual-resolution scheme. First, a low-resolution depth network G_D^l estimates coarse normal maps N_l and depth maps D_l from a down-sampled version of I. Though lacking fine details, this module captures the global geometry reliably and serves as a strong prior.

Then, a high-resolution depth network G_D^h, composed of shallow cascaded CNNs, refines these coarse maps into full-resolution double-sided depth maps D_h. It fuses N_h and D_l to produce geometrically consistent high-frequency details, avoiding the artifacts common in direct depth regression from RGB.

3.1.3. Mesh Generation

The depth maps D_h are converted into 3D point clouds using perspective projection. We then apply a screened Poisson surface reconstruction to generate the final mesh M, which captures detailed surface geometry including folds, accessories, and extremities. Compared to implicit-based methods, this explicit mesh generation is efficient and resolution-agnostic, supporting fine surface topology.

3.2. Full-body Texture Prediction and Refinement via SIFU

Although 2K2K produces detailed geometry, the output mesh lacks color and texture information. We address this by introducing a second stage that predicts high-quality, full-body textures using an enhanced version of the SIFU pipeline. This stage comprises two key modules.

3.2.1. Side-view Conditioned Implicit Function

We first estimate the corresponding SMPL-X model from mesh M to serve as a geometric prior. Side-view normals (left, right, back) are rendered from the SMPL-X mesh and used as queries in a Side-view Decoupling Transformer. The transformer performs cross-attention with global image features to extract side-specific features that are difficult to observe directly in monocular input.

We then apply a hybrid prior fusion strategy to aggregate both pixel-aligned and mesh-projected features at each 3D query point. Specifically, we combine feature descriptors from front and side planes, spatially aligned using barycentric interpolation over SMPL-X vertices. These fused features, along with signed distance fields and normal embeddings, are passed through an implicit function $\mathrm{MLP}(\cdot)$ to predict both occupancy and RGB color at each point:

$$(o, c) = \mathrm{MLP}(F_S(x), F_P(x), \mathrm{SDF}(x), F_N(x)). \quad (1)$$

The coarse textured mesh M_c is then extracted using Marching Cubes, with per-vertex RGB color predicted from the implicit decoder.

3.2.2. Mesh Simplification

Due to the complexity of the 2K2K mesh, which results in long mesh mapping times, we perform mesh simplification on the original mesh. We use the simplification algorithm provided by PyMeshLab, specifically the quad edge collapse technique. This method reduces the number of vertices in the mesh by selectively merging the edges, thereby simplifying the mesh's geometric structure. Specifically, the algorithm progressively merges neighboring edges and faces while ensuring that the original shape and details are maintained, thus reducing redundant geometric data. This

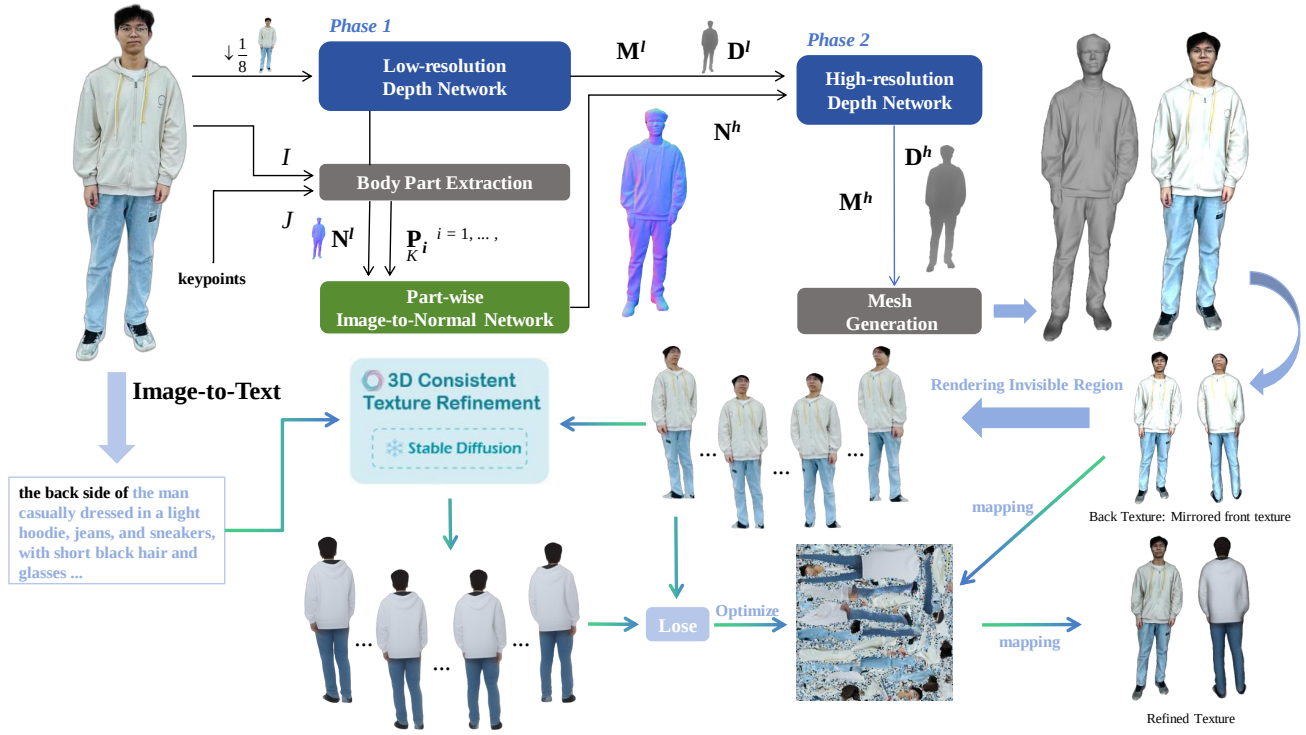

Figure 3. An overall framework of the proposed method. The first stage constructs a high-fidelity mesh from a single image, while the second stage performs full-body clothed texture prediction and refinement. The two-stage framework outputs a high-fidelity human mesh with full-body precise texture map.

simplification process not only effectively reduces computational complexity, but also improves the efficiency of subsequent processing and rendering.

3.2.3. Image to Text Conversion

We use image-to-text models to create descriptive text prompts that facilitate the diffusion-based texture enhancement process. Specifically, we use the advanced capabilities of GPT-4v to generate precise descriptions of the input image. After uploading an image, we prompt the model with, "lease describe the person in the image in English, focusing on their clothing, colors, style, and hairstyle, without including the background. Limit the description to under 70 words." The generated description is represented as A. To further customize the prompt, we prepend phrases like "the back side of A, realistic, vivid" to create a comprehensive input prompt P.

3.2.4. 3D Consistent Texture Refinement with Diffusion Prior

To improve realism and fill in occluded regions, we perform texture refinement using a pre-trained text-to-image diffusion model. We first generate a natural language prompt P from the input image using a vision-to-text module, describing the person's clothing and appearance.

We render the mesh M_c from multiple novel views I_k, and refine these using diffusion-based semantic editing, ensuring 3D consistency across views. This is achieved via DDIM inversion and consistent token propagation across layers. The refined images J_k are projected back to update the UV texture map T of the mesh.

The texture is optimized using a weighted combination of photometric MSE loss, VGG perceptual loss, Chamfer loss (on rendered mesh), and a fidelity loss to the original view:

$$\min_{T} \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{VGG}} + \lambda_3 \mathcal{L}_{\text{CD}} + \lambda_4 \mathcal{L}_{\text{fidelity}}. \quad (2)$$

This process ensures the resulting texture is high-resolution, style-consistent with the input, and semantically complete even in regions not visible from the camera view.

Our two-stage pipeline effectively decouples the 3D reconstruction and texturing tasks. By combining part-wise geometric prediction from 2K2K with visibility-aware texture refinement in SIFU, our framework achieves superior results in both geometry and appearance. It is robust to pose and clothing variation, supports arbitrary single-view images, and scales efficiently for real-world deployment.

4. Experiments

We selected some examples as a dataset for evaluation and compared the results with the 2k2k.

4.1. Evaluation

4.1.1. Metrics

The texture optimization quality of our model is quantitatively evaluated using PSNR, SSIM, and LPIPS, by comparing the multi-view images rendered from the optimized mesh with the ground truth.

PSNR (Peak Signal-to-Noise Ratio): PSNR is a commonly used metric for assessing the quality of reconstructed images. It compares the pixel-by-pixel difference between the original and the reconstructed image, measuring the ratio between the maximum possible power of an image and the noise affecting its quality. Higher PSNR values indicate better image quality, with fewer distortions and noise.

SSIM (Structural Similarity Index Measure): SSIM is a perceptual metric that measures the structural similarity between two images. Unlike pixel-based metrics like PSNR, SSIM considers changes in structural information, luminance, and texture. It compares the perceived quality by assessing how similar the local patterns of pixel intensities are in terms of luminance, contrast, and structure. A higher SSIM value suggests that the reconstructed image is closer to the original in terms of visual perception.

LPIPS (Learned Perceptual Image Patch Similarity): LPIPS is a deep learning-based perceptual similarity metric that evaluates the perceptual differences between two images. It uses pre-trained neural networks to capture high-level features, comparing them between the original and generated images. LPIPS tends to align more closely with human perception, and lower values suggest that the two images are visually similar, with minimal perceptual differences.

4.1.2. Quantitative Evaluation

In terms of texture reconstruction, our model performs well, as shown in Tab. 1. The average PSNR value of the multi-view images rendered by the model is **23.028**, indicating a medium level of image quality. The SSIM value is **0.908**, indicating that the rendered images are structurally similar to the original images, suggesting that the model performs well in terms of structural details of the texture. The LPIPS value is **0.1216**, indicating that the rendered images are very close to the original images, with minimal perceptual differences.

4.1.3. Qualitative Results

Our results demonstrate the model's excellent performance on images with various poses. As shown in Fig. 4, our model is capable of handling complex scenes such as loose clothing and challenging poses.

Figure 4. Qualitative results of our method. Our approach achieves high-fidelity human mesh construction with full-body detailed textures, with a single image input.

4.2. Applications

Through this section, we demonstrate the model's versatility and potential in digital workflows, as well as its broad application prospects across various industries.

Texture Editing. With the powerful capabilities of text-to-image diffusion models, we can easily generate edited textures in 3D consistent texture optimization by changing the text prompts.

Scene Building and 3D Printing. The model's precise geometry and optimized textures make it highly suitable for virtual scene creation and 3D printing. It enhances realism in simulations and games, while simplifying the 3D printing process and reducing the need for complex scanning. This has potential applications in rapid prototyping, educational resources, and customized 3D sculptures.

5. Conclusion

In this paper, we present a two-stage framework for high-fidelity clothed 3D human reconstruction from a single image. Our approach leverages the strengths of two complementary systems: 2K2K for accurate and detailed mesh generation, and SIFU for photorealistic texture prediction and refinement. By decoupling geometry and appearance modeling, we effectively address the challenges of reconstructing full body mesh with high-fidelity textures from single input.

Through part-wise normal estimation and multi-resolution depth refinement, the 2K2K module yields high-quality meshes with rich geometric detail. The subsequent texture generation stage builds upon this geometry using a side-view conditioned implicit function and a 3D-consistent diffusion-based refinement strategy. This results in realistic, full-body textures that are both view-consistent and semantically aligned with the input image.

Extensive qualitative and quantitative evaluations on

Dataset	PSNR	SSIM	LPIPS
2k2k front (12 persons average)	25.438	0.947	0.056
2k2k back (12 persons average)	18.449	0.878	0.168
2k2k total average	21.944	0.913	0.112
Ours front (12 persons average)	26.332	0.951	0.058
Ours back (12 persons average)	19.724	0.866	0.185
Ours total average	23.028	0.908	0.121

Table 1. Quantitative evaluation of texture optimization quality of our model in comparison with 2k2k baseline.

public benchmarks demonstrate that our method achieves great performance in both mesh accuracy and texture realism. Furthermore, our framework is robust to pose variation and generalizes well to in-the-wild images, making it suitable for downstream applications such as AR/VR, character animation, and 3D printing.

In future work, we plan to further enhance the realism and efficiency of the pipeline by integrating more lightweight transformer backbones, incorporating temporal consistency for video-based reconstruction, and exploring user-controllable text-driven texture editing in real time.

References

[1] Badour Albahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, page 1–11. ACM, 2023. 3

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing, 2022. 3

[4] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE/CVF on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[5] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K. Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction, 2022. 3

[6] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction, 2023. 3

[7] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 42:2523–2539, 2015. 1

[8] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing controllable avatars, 2023. 3

[9] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13, 2016. 1

[10] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes, 2023. 4

[11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition, 2023. 4

[12] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images, 2023. 2, 3

[13] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[14] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction, 2020. 3, 4

[15] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[16] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited, 2022. 3

[17] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[18] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans, 2023. 3

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed hu-

mans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[20] Sai Sagar Jinka, Rohan Chacko, Avinash Sharma, and PJ Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 3

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

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

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[24] 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)*, 34:1–16, 2015. 3

[25] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, 1987. 2

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 2

[27] 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 of the IEEE/CVF on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[28] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018. 2

[30] 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/CVF on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[31] Marco Pesavento, Marco Volino, and Adrian Hilton. Super-resolution 3d human shape from a single low-resolution image. *arXiv preprint arXiv:2208.10738*, 2022. 2

[32] Shenhan Qian, Jiale Xu, Ziwei Liu, Liqian Ma, and Shenghua Gao. Unif: United neural implicit functions for clothed human reconstruction and animation, 2022. 2, 3

[33] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[34] RealityCapture. https://www.capturingreality.com/. 2

[35] 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 (ICCV)*, 2019. 2, 3

[36] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[37] Twindom. https://web.twindom.com/. 2

[38] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[39] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[40] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[41] 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 (TPAMI)*, 44:3170–3184, 2021. 2