Single-view 3D Human Reconstruction Using Generative Prior

Zhengmao Liu
Stanford University
450 Jane Stanford Way
zliu1019@stanford.edu

Abstract

We present a novel approach to reconstruct complete 3D clothed human models from single front-view images. We address the challenges of realistic textures and accurate poses by: (1) estimating the 3D human body model; (2) generating normal maps for unseen viewpoints (side and back) and creating coarse textured mesh; and (3) refining the clothes textures of the unseen views. Initially, coarse textures are generated by fusing side-view normal maps with the input images. These textures are then refined using a diffusion model, guided by the 3D human pose from SMPL-X.
The paper is organized as follows: Section 1 provides background information and an overview of our proposed approach. Section 2 reviews relevant literature on 3D human modeling, pose generation with diffusion models, and texture consistency. Section 3 details our methodology and architectural design. Sections 4 presents our dataset and experimental results. Section 5 outlines future improvement and research directions. Section 6 concludes the project work. Lastly, Section 7 contains the acknowledgments.

1. Introduction

The reconstruction of dressed 3D human figures presents significant challenges but offers substantial benefits across domains including Augmented Reality (AR), Virtual Reality (VR), gaming, film-making, fashion, e-commerce, healthcare, and education. This project addresses these challenges by focusing on generating 3D human models from single-view images, such as selfies and online photos.

Traditional 3D human reconstruction techniques generally rely on multi-view images to achieve comprehensive scene observability, ensuring accurate and detailed reconstruction. However, this requirement presents a significant challenge in practical applications. Most real-world image sources, such as those captured by mobile devices or shared on the internet, typically provide only single-view perspectives (e.g., front or side views). This limitation restricts the ability to obtain the necessary multiple viewpoints, thereby impeding the effectiveness and applicability of conventional 3D reconstruction methods in everyday scenarios.

Our project takes a single front-view image of one person as input to generate a 3D human pose model and a coarse mesh. We begin by predicting the 3D human pose in the SMPLX [8] representation using the PIXIE [3] model. PIXIE integrates expert sub-networks for body, face/head, and hand regression within a unified framework. Then, we generate a coarse mesh by fusing normal maps, derived from the SMPLX model, with the front-view image. This is achieved using the cross-attention mechanism and a multilayer perceptron architecture, similar to the approach described in SIFU [12]. To refine the clothing textures on the side and back views, we fine-tuned a generative model using CHAMP [13], where we use the 3d human pose model as a guide to incorporate visual details on the side and back views.

In summary, this project integrates SMPLX's comprehensive 3D human representation with SIFU's single-view coarse mesh generation capabilities. Furthermore, we utilize CHAMP's controllable Latent Diffusion Model (LDM) to ensure texture and pose consistency.

2. Related Work

Conventional methods for 3D human reconstruction [6, 2] can develop detailed and finely crafted 3D human models. However, these approaches are limited in their applicability to internet or phone images due to constraints on available viewing angles.

PIFU [9] achieves promising results by employing implicit functions that require only single-view images to predict 3D human models and generate novel view syntheses. Despite its effectiveness, PIFU's dependency on 3D ground truth data and normal maps restricts its use in complex real-world scenarios. Moreover, it struggles with depth ambiguity, which can result in unnaturally elongated features in the reconstructed human meshes.

SMPL [7], which provides skinned 3D body models designed for various genders and a wide range of poses, is a robust framework for capturing human motion. It represents the 3D human body using parameters that include shape ($\beta \in \mathbb{R}^{10\times1}$), 24 joints and their relative rotation poses (($\theta \in \mathbb{R}^{24\times3}$). SMPL-X [8] extends SMPL by adding parameters for hands, jaws, facial expressions, and poses. Despite its increased complexity, SMPL-X offers significant advantages for applications that require high levels of expressiveness and detailed representation of the human face and hands.

SHERF [5] employs a single-view image in conjunction with the SMPL model [7] to reconstruct 3D models of clothed humans and enable novel view synthesis.This process leverages neural radiance fields, which allow for producing detailed and realistic representations of human figures. Although this regression-based method demonstrates effectiveness in interpolation, enabling the generation of intermediate views between known angles, it faces certain limitations, particularly when it comes to hallucination. This issue arises when the system attempts to infer novels views, leading to potential inaccuracies and inconsistencies in the reconstructed geometry and appearance. Additionally, SMPL models struggle to accurately depict detailed facial features and hands. These challenges highlight the need for further advancements to improve the model's capability in handling diverse viewing angles and enhancing the overall realism of clothing textures.

SIFU [12] has the unique advantage of reconstructing 3D models of clothed humans from only front-view images. The process begins with estimating the SMPL-X [8] human pose model from the front-view image, which serves as the basis for creating a coarse 3D human model. To refine textures, SIFU employs a diffusion model guided by text descriptions. Initially, it generates a text description based on the front-view image, and then manually incorporates textual constraints to refine textures for side and back views within the diffusion model. However, this approach has limitations—since the diffusion model does not incorporate constraints from the front-view images during refinement, the resulting textures can significantly differ from the original front-view image, relying solely on text-based constraints.

CHAMP [13] is per-frame probabilistic generative model for video generation. It utilizes the SMPL [7] model as the 3D human parametric model to establish a unified representation of body shape and pose. This facilitates the accurate capture of intricate human geometry and motion characteristics from source videos. Specifically, it incorporates rendered depth images, normal maps, and semantic maps obtained from SMPL sequences, alongside skeleton-based motion guidance, to enrich the conditions to the latent diffusion model with comprehensive 3D shape and detailed pose attributes. This method facilitates video generation with consistent texture and poses. However, it lacks temporal context leads to generation of details that are irrelevant to past observations. In addition, it doesn't have comprehensive representation of faces and hands using SMPL models, which might incapable of presenting complex motions.

MagicMan [4], which uses a single view image as input to generate novel view synthesis, is a 3D-aware diffusion model and it uses the iterative refinement pipeline to ensure the quality and consistency of generated views. However, this work also has the limitations of low quality textures of human faces and hands due to the SMPL 3d human representation.

3. Method

We divide the construction of the fine clothed 3D human model into three stages to achieve the desired results. The overall pipeline is in Fig 2.

First, we utilize the front-view image as input to predict the 3D SMPL-X [8] body representation using the PIXIE [3] model. The PIXIE model employs a suite of specialized sub-networks dedicated to body, face/head, and hand regression, which are subsequently integrated into a more comprehensive network to generate the 3D human body (see Fig. 3).

Next, we utilized side-view decoupling transformers and the Hybrid Prior Fusion Strategy as described in SIFU [12]. The side-view decoupling transformer processes front-view image features, denoted as $F_{front} \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$, and embeds them into the encoder. Meanwhile, normal maps from the side and back views are embedded and used as queries in the transformer decoder. Cross attention is applied using the following equation:

$$f(z_i, h) = Softmax(\frac{(z_i W^Q)(hW^K)^T}{\sqrt{d}})(hW^V) \qquad (1)$$

Here, z_i represents the embeddings from the side views (*left, back, right*), and h is the embedding from the front-view images. The resulting outputs from each side and back view are combined with the front view encoder output to generate a coarse mesh using the hybrid fusion strategy.

The hybrid fusion technique [11] efficiently merges features at a query point by leveraging spatial localization

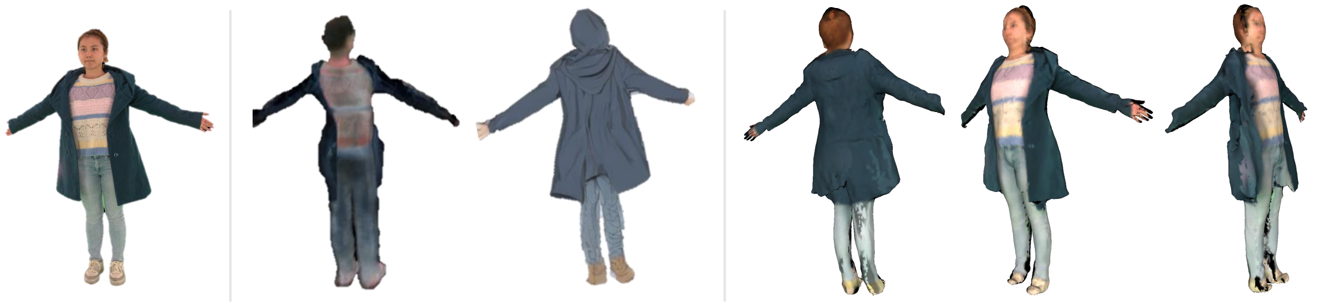

Figure 1: Using the single view input, we generated novel view images with SHERF, SIFU, and our proposed method. The results clearly highlight that our approach produces a back view that is far more realistic and consistent with the front view textures and poses than the other methods.

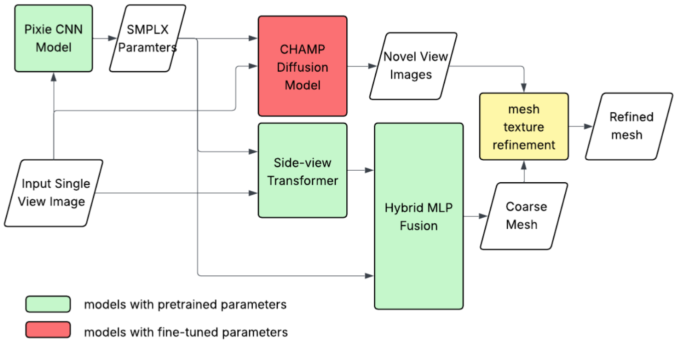

Figure 2: Pipeline to generate the refined mesh from the single view Images

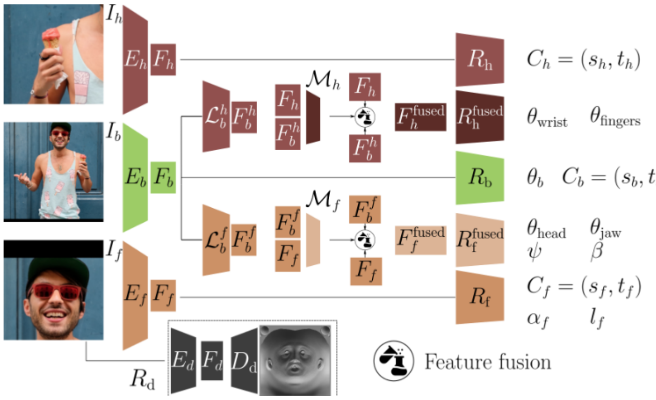

Figure 3: PIXIE[3] architecture , the body, face/head and hand images are fed to the expert encoder to produce part-specific features. These features are fused in M_f, M_h to get the parameters of head, body, face

.

and human body prior knowledge. The feature maps F_j (where $j \in front, left, back, right$) from the side-view transformer are divided into two groups. In the first group, the feature at a query point x is obtained by projecting the point onto the feature map and integrating information from the four views as follows:

$$F^S(x) = F^S_f(x) \oplus avg(F^S_l(x), F^S_r(x), F^S_b(x)) \quad (2)$$

In the second group, the SMPL-X [8] mesh is employed to obtain vertex information from the query point x using barycentric interpolation:

$$F^P(x) = uF^S(v_0) + vF^S(v_1) + wF^S(v_2) \quad (3)$$

Here, $F^S(v)$ is achieved by projecting SMPL-X [8] mesh vertices onto the feature maps. $[v_0, v_1, v_2]$ represent the nearest triangular face to the query point x, and $[u, v, w]$ are the barycentric coordinates of the query point x projected onto the triangle formed by $[v_0, t_1, v_2]$.

The features obtained from these queries are concatenated along with the SMPLX-mesh $SDF(x)$ and normal features, which are then input into a multilayer perceptron network to predict occupancy and color, thereby generating a coarse mesh. The architecture of the side-view transformer and hybrid fusion strategy is in: Fig 4.

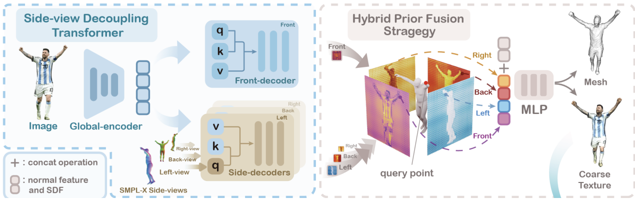

Figure 4: SIFU[12] coarse mesh generation Architecture of the coarse mesh generation in SIFU

.

To enhance the mesh's texture, we utilized CHAMP [13], a per-frame probabilistic generative model, the architecture is in Fig 5. This involves creating novel side and back view images based on the initial front view and the reference SMPL parameters. we fine-tune the Stable Diffusion network $U_{denoiser}$ and a reference network U_{ref} where the U_{ref} provides the appearance information of the input image and the $U_{denoiser}$ renders novel views conditioned on the appearance information. The texture refinement is then achieved by back-projecting the pixel information from these generated views onto the coarse mesh.

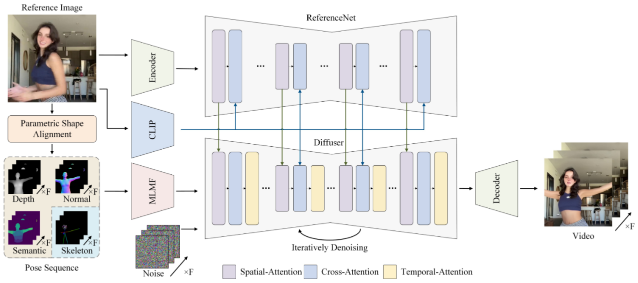

Figure 5: CHAMP[13] architecture for novel view generation.

.

4. Experiments

4.1. Dataset

We use 4D-Dress [10] datasets in this project. The 4D-Dress dataset is a multi-view video dataset that consists of 64 subjects wearing a variety of real-world outfits, including challenging loose garments and jackets. Each subject performs eight different motions, captured by four uniformly distributed cameras. Each motion sequence contains roughly 200 frames.

4.2. Qualitative Results

We initially employed SHERF [5] for novel view synthesis, as demonstrated in Fig. 6. While SHERF effectively captured the overall human pose and the textures of the sweaters and pants, it struggled to accurately reconstruct the details of the jackets and faces. This led to inconsistencies between the generated side and back views and the front view.

Following the SIFU process, including its refinement model, we also evaluated the generated novel views and textured mesh. A major drawback was the reliance on a diffusion model without image guidance. This resulted in refined novel views and mesh that appeared less realistic and exhibited significant discrepancies compared to the input images (See Fig 6).

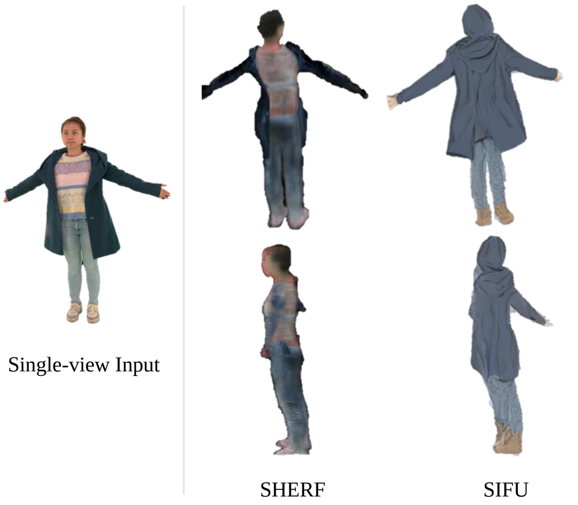

Figure 6: Experiment results using SHERF and SIFU

In this project, we apply CHAMP to get consistent pose and textures for novel views. This is achieved by fine tuning the Stable Diffusion network $U_{denoiser}$ and a reference network U_{ref} where the U_{ref} provides the appearance information of the input image and the $U_{denoiser}$ renders novel views conditioned on the appearance information. The novel view images are then projected to the corresponding vertices to obtain the fine textured meshes of the 3d clothed human (Fig 7).

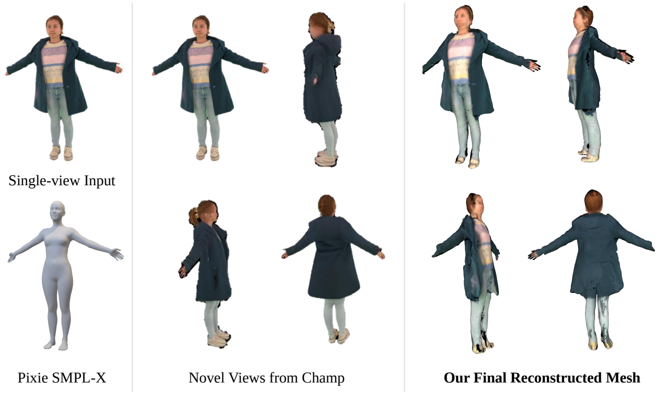

Figure 7: Result from our proposed approach

5. Future work

Comprehensive Evaluation. Due to time constraints, we could not conduct a thorough quantitative evaluation in this version. In future work, we plan to evaluate our reconstructed meshes using standard geometric metrics such as the Chamfer Distance and the Earth Mover's Distance (EMD) to measure the accuracy of 3D shape reconstruction against ground-truth meshes. To assess rendering quality, we will report image-based metrics including PSNR and SSIM. In addition, we will incorporate perceptual metrics such as LPIPS and FID to better capture human perceptual alignment and overall visual quality.

Texture Improvement. While our method captures the overall geometry accurately, there is still room for improving texture quality. Currently, we compute the final pixel colors by performing a weighted average of sampled input pixels, using visibility as the weighting factor. A more sophisticated approach would incorporate both viewing direction and surface normal to better resolve view-dependent appearance. This could lead to sharper and more consistent textures, particularly in regions with complex lighting or self-occlusion.

Model Fine-tuning and Architecture Enhancement. In the current pipeline, we only fine-tuned the CHAMP model to obtain detailed texture information. Future work could involve fine-tuning earlier stages, such as the side-view transformers and the hybrid fusion MLP network. It is also worth noting that our side-view transformer utilizes the ViT architecture, which may be inefficient and might miss some fine-grained details due to its patch-based embedding approach. Exploring more advanced Vision Transformer architectures like DETR [1] or alternative CNN-based embeddings could improve feature extraction. Lastly, incorporating temporal information could further enhance the accuracy of the 3D human models.

6. Conclusion

This paper presents a novel pipeline for generating 3D models of clothed humans from single-view (front) images. Initially, we use PIXIE to generate the 3D SMPLX parameters. These parameters, along with the front-view images, are fed into pretrained side-view transformers and a hybrid fusion MLP network to produce a coarse 3D mesh. Additionally, the front-view and SMPLX parameters are utilized in the fine-tuned CHAMP diffusion model, trained on the dress4d dataset, to create back and side view images. These images are then projected onto the coarse 3D mesh to obtain a finely textured 3D model. The results, as illustrated in Figure 7, demonstrate superior performance compared to state-of-the-art methods like SHERF and SIFU, particularly in areas such as facial expressions, hand poses, texture quality, and pose consistency.

7. Acknowledgment

I would like to extend special thanks to Dr. YoungJoong Kwon for her invaluable guidance and assistance with the dataset and fine-tuning of the CHAMP model. Her support was instrumental in achieving the impressive results of this project.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers, 2020.

[2] J. Chen, W. Yi, T. Wang, X. Li, L. Ma, Y. Fan, and H. Lu. Pixel2isdf: Implicit signed distance fields based human body model from multi-view and multi-pose images, 2022.

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

[4] X. He, X. Li, D. Kang, J. Ye, C. Zhang, L. Chen, X. Gao, H. Zhang, Z. Wu, and H. Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement, 2024.

[5] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu. SHERF: Generalizable human nerf from a single image, 2023.

[6] Z. Li, M. Oskarsson, and A. Heyden. Detailed 3d human body reconstruction from multi-view images combining voxel super-resolution and learned implicit representation, 2020.

[7] G. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graphics*, 34(6):248:1–248:16, 2015.

[8] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CoRR*, abs/1904.05866, 2019.

[9] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[10] W. Wang, H.-I. Ho, C. Guo, B. Rong, A. Grigorev, J. Song, J. J. Zarate, and O. Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[11] Z. Zhang, L. Sun, Z. Yang, L. Chen, and Y. Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction, 2023.

[12] Z. Zhang, Z. Yang, and Y. Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction, 2024.

[13] S. Zhu, J. L. Chen, Z. Dai, Q. Su, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance, 2024.