Building An AI-Powered Fashion Application: Virtual Clothing Try-On

Shawn Zhang
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
shawn.zhang@stanford.edu

Abstract

This project explores a full pipeline for virtual try-on and 3D human reconstruction, integrating multiple state-of-the-art methods across 2D synthesis, 3D modeling, and rendering. I begin by using VITON-HD to generate high-resolution (1024×768) 2D try-on images, which tackle limitations in previous virtual try-on systems by introducing ALIAS normalization and generator architectures for more realistic image synthesis. Next, I apply the ECON pipeline to generate a detailed 3D human mesh from a single image using normal map estimation and geometry inpainting techniques. I address file structure and compatibility challenges for integrating ECON outputs with CEB_ECON, a Blender-based tool for visualization and texture projection.

Despite successful mesh generation and partial setup for texture rendering, my pipeline encountered challenges in projecting 2D clothing textures onto the 3D mesh due to unresolved UV mapping issues. This report details each technical component, installation and debugging workflows, and reflections on model strengths, integration gaps, and possible improvements.

1. Introduction

Virtual try-on and 3D human reconstruction are two rapidly evolving tasks in computer vision that address real-world needs in e-commerce, gaming, augmented reality (AR), and virtual content creation. In particular, image-based virtual try-on allows users to visualize how different clothing items might look on their own body, without physically trying them on—an application that has gained increasing relevance in the digital retail industry. Similarly, reconstructing animatable 3D human avatars from 2D images plays a critical role in virtual reality and film production, enabling lifelike renderings and character control from minimal input.

The motivation behind this project is twofold: first, to explore the technical feasibility of high-resolution, realistic virtual try-on experiences from a single image; and second, to investigate how to reconstruct textured, animatable 3D models from those same inputs. While these tasks are individually challenging, integrating them into a single pipeline presents practical difficulties in interoperability, resolution scaling, and file structure consistency—making the end-to-end system design both technically rich and practically valuable.

The input to the pipeline consists of a single RGB image of a person (front-facing) and a target clothing image. First, using a deep generative adversarial network (GAN) architecture (VITON-HD), the system overlays the clothing item onto the image, producing a photo-realistic try-on image of resolution 1024×768. Next, the system reconstructs a clothed 3D human mesh from the same input image using a hybrid implicit-explicit representation (ECON), which internally relies on normal map estimation, SMPL-X fitting, and geometry inpainting. Finally, the pipeline attempts to project 2D textures onto the reconstructed 3D model using a Blender add-on (CEB_ECON), enabling visual inspection and animation within Blender.

This project focuses not only on deep learning-based model inference, but also on system integration, debugging, and visualization—highlighting the importance of tools and infrastructure in deploying computer vision models for real-world, interactive applications.

2. Related Work

This project lies at the intersection of three active areas in computer vision: virtual try-on, 3D human reconstruction, and texture mapping for animatable avatars. Each field has developed independently but shares overlapping technical foundations, such as human parsing, image-to-image translation, and generative modeling.

Current state-of-the-art systems address the subproblems in isolation. Virtual try-on pipelines produce high-res 2D imagery but lack physical consistency. 3D reconstructions are geometrically detailed but lack texture. Texture methods are advancing, yet depend on fragile integration pipelines. This project attempts to integrate multiple SOTA tools into a unified workflow, revealing practical barriers in repro-

ducibility and automation.

2.1. Virtual Try-On Systems

Early virtual try-on methods, such as VITON [4], pioneered the task by using a two-stage approach: generating human parsing maps and warping garments using thin-plate spline transformations. However, these models were constrained by low resolution (e.g., 256×192) and suffered from texture blurriness. CP-VTON [9] improved garment alignment using dense pose estimation, while ClothFlow [3] introduced flow-based warping for better geometric realism.

The state-of-the-art VITON-HD [2] significantly increases the output resolution to 1024×768 by introducing ALIAS normalization, which improves alignment of warped garments with target body regions. While highly effective in preserving clothing details, VITON-HD still relies on 2D overlays, which limits its ability to support 3D interactivity or reanimation.

2.2. Human 3D Reconstruction

Reconstructing 3D humans from a single RGB image is an under-determined problem. Parametric body models like SMPL [5] and SMPL-X [6] offer pose and shape priors to regularize this task, but they often fail to capture clothing and hair details. PIFu [7] introduced pixel-aligned implicit functions to represent detailed surfaces but struggles with occlusions. ICON [10] improved on this by using normal maps and silhouette consistency for higher-fidelity clothed reconstructions.

The ECON model [11] builds on ICON by combining implicit functions with normal-based surface inference and explicit priors from SMPL-X, allowing for more accurate geometry reconstruction, even with loose clothing or non-canonical poses. However, ECON outputs require manual intervention to fit into downstream pipelines, such as animation or texture projection.

2.3. Texture Mapping and Animation

Once 3D geometry is reconstructed, mapping realistic textures is critical for photorealism. Techniques like Neural Textures [8] and TEXTure [1] synthesize detailed textures from partial observations using GANs. TEXTure, used in this project, projects 2D views onto 3D meshes via differentiable rendering, but requires manual file structure setup and lacks robust automation. Blender-based add-ons like CEB_ECON bridge the gap between research outputs and visualization, though they depend heavily on user input for correct pathing and folder hierarchies.

3. Methods

This project integrates multiple deep learning models and software pipelines to achieve high-resolution, image-based 3D virtual try-on. The overall method can be decomposed into three major stages: (1) 2D virtual try-on using VITON-HD, (2) 3D human reconstruction using ECON, and (3) texture projection and visualization using CEB_ECON with Blender. Below we describe the algorithms, architectures, and mathematical formulations used in each component.

3.1. VITON-HD: High-Resolution 2D Virtual Try-On

The input to the VITON-HD model is a pair of images: a person image $I_p \in \mathbb{R}^{3\times H\times W}$ and a target clothing image $I_c \in \mathbb{R}^{3\times H\times W}$, both resized to 1024×768. The output is a synthesized image $I_{tryon} \in \mathbb{R}^{3\times H\times W}$, where the target clothing is photo-realistically aligned and blended onto the person.

The pipeline consists of three components:

1. Segmentation Generator: A semantic parser generates a segmentation map $S_p \in \mathbb{R}^{K\times H\times W}$ of the person.

2. Geometric Matching Module (GMM): A Thin-Plate Spline transformation aligns I_c with the body pose and segmentation using keypoint regression. It outputs a warped clothing image I'_c.

3. ALIAS Generator: The core generator applies Alignment-Aware Segment normalization to adaptively modulate features based on segmentation classes and refine the final try-on image.

The generator is trained using a combination of pixel-wise L1 loss, perceptual loss $\mathcal{L}perc$, and adversarial loss $\mathcal{L}adv$. The total loss is:

$$\mathcal{L}total = \lambda_1\mathcal{L}L1 + \lambda_2\mathcal{L}perc + \lambda_3\mathcal{L}adv \quad (1)$$

3.2. ECON: High-Fidelity 3D Human Reconstruction

The input to ECON is a single RGB image $I_p \in \mathbb{R}^{3\times H\times W}$. The goal is to reconstruct a detailed and animatable 3D human mesh M consisting of explicit geometry and optionally enhanced body parts.

The pipeline proceeds in three stages:

1. Normal Map Estimation: Two convolutional networks generate front and back normal maps $N_f, N_b \in \mathbb{R}^{3\times H\times W}$, representing detailed surface normals.

2. Depth Reconstruction: The normal maps are converted into 2.5D depth maps D_f, D_b, which are lifted into partial meshes using point cloud estimation and surface fitting.

3. Surface Fusion: The front and back meshes are merged using Poisson surface reconstruction and refined using

IFNet, a continuous implicit surface decoder trained to fill in missing geometry.

Optionally, the reconstructed body mesh can be augmented with high-quality hands and face from SMPL-X to improve animation fidelity. The reconstruction is supervised by geometric consistency loss:

$$\mathcal{L}geo = \|N_{pred} - N_{gt}\|2 + \|D_{pred} - D_{gt}\|2 \quad (2)$$

where N_{gt} , D_{gt} are ground-truth normals and depth from synthetic datasets.

3.3. CEB_ECON + TEXTure: Texture Projection in Blender

To transfer the visual appearance from the input image onto the 3D model, we use the CEB_ECON Blender plugin, which wraps ECON and the TEXTure diffusion model. The input to this pipeline is the reconstructed mesh M and the original image I_p. The output is a fully textured 3D model rendered in Blender.

Steps include:

1. Mesh Preparation: The .obj output from ECON is renamed and moved to TEXTure/experiments/input_0/mesh/mesh.obj to satisfy TEXTure's loading conventions.

2. Texture Diffusion: TEXTure uses a conditional diffusion model to generate view-consistent texture maps $T \in \mathbb{R}^{3\times H_t \times W_t}$, conditioned on the geometry and the input image.

3. Projection & Rendering: Blender's material node system is used to attach the texture map T to the UV-mapped mesh M, enabling real-time visualization.

The underlying model used for TEXTure is a latent diffusion model (LDM) that optimizes the following objective:

$$\mathcal{L}ldm = \mathbb{E}x_t, \epsilon \sim \mathcal{N}(0,1)\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right] \quad (3)$$

where x_t is the noisy image at timestep t, ϵ_θ is the denoiser, and c encodes the 3D mesh condition.

3.4. Integration and Modifications

Each component of the pipeline was adapted from public repositories:

- VITON-HD from shadow2496/VITON-HD. (https://github.com/shadow2496/VITON-HD)
- ECON from ECON. (https://xiuyuliang.cn/econ/)
- CEB_ECON and TEXTure from kwan3854/CEB_ECON and TEXTurePaper/TEXTure. (https://github.com/kwan3854/CEB_ECON?tab=readme-ov-file)

Custom integration scripts and manual fixes were added to:

- Align file structures and naming conventions.
- Work around broken automated downloads.
- Address dependency issues (e.g., mediapipe, cl.exe, PyTorch3D on CPU).
- Manually assign texture images in Blender's Shader Editor due to default failures in automatic projection.

4. Dataset and Features

This project leverages a combination of publicly available datasets and pre-trained models for virtual try-on and 3D human reconstruction. Specifically, the pipeline integrates person images and clothing items to generate 2D try-on results and reconstruct 3D clothed human meshes.

4.1. Data Sources

1. VITON-HD Dataset: For the 2D virtual try-on component, I used the VITON-HD dataset, which includes high-resolution person images and paired clothing item images. The dataset contains:

   - 11,647 image pairs in the training set.
   - 2,032 image pairs in the test set.

   Each pair consists of a front-view image of a person and an upper-body clothing item, both with a resolution of 1024×768 pixels.

2. DressCode Dataset: For clothing diversity and conditional generation, I plan to use more sample assets from the DressCode dataset, which includes multiple human images with detailed garment annotations across casual, formal, and sport outfits. It provides:

   - Over 50,000 images across multiple styles.
   - SMPL-based ground truth meshes.
   - Semantic segmentation maps and pose keypoints.

3. Input Image for ECON 3D Reconstruction: I used a single-view image of a person (e.g., input_0.jpg) with resolution 1024 × 768 as the input to ECON's 3D reconstruction pipeline. These images are similar in format to those used in VITON-HD and were manually selected for visual clarity and pose alignment.

4.2. Features and Preprocessing

- Segmentation Maps: VITON-HD relies on segmentation maps generated from the input image to localize body parts and clothing regions. These maps are extracted using a U-Net-like architecture and are used as spatial guidance for the ALIAS Generator.

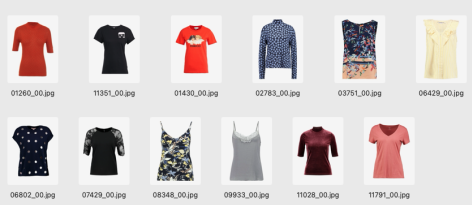

Figure 1. Clothing Items

- Pose Estimation DensePose: Body pose is estimated using DensePose to provide coordinate-based conditioning. These pose maps are particularly important in warping the clothing to match the body configuration.
- Normal Maps (ECON): For 3D model reconstruction, ECON infers high-fidelity front and back normal maps from the single input image. These normal maps serve as geometric cues that are converted into partial 2.5D surfaces, which are later inpainted to produce full 3D geometry.
- SMPL-X Meshes: ECON further integrates SMPL-X body models to align and complete the geometry of the reconstructed humans. This involves canonical mesh parameterization using estimated shape and pose parameters.
- Clothing Textures: While the pipeline includes steps to project textures from 2D images onto the 3D mesh using CEB_ECON and TEXTure, texture mapping encountered integration issues in this implementation and was not fully successful.

4.3. Sample Data

Figures 1 and 2 are data used in this project, which is sourced from the VITON-HD dataset and designed for the task of high-resolution virtual try-on synthesis. It contains two primary types of images: Figure (1) includes images of clothing items with transparent backgrounds, and Figure (2) are images of fashion models wearing different clothing. These two sets of data are used as inputs for conditioning and reference, respectively, in the virtual try-on pipeline.

5. Experiments / Results / Discussion

All experiments were conducted on a local Windows machine without access to an NVIDIA GPU. Since ECON is designed to run using CUDA acceleration, I manually edited its source code to bypass CUDA dependencies, enabling CPU-only inference. While this significantly increased computation time (especially during normal map generation and mesh fusion), it confirmed that ECON can function without a GPU, albeit at a performance tradeoff.

To ensure compatibility, I resolved missing dependencies such as torchvision, trimesh, and Microsoft

Figure 2. Fashion Models Wearing Original clothing

Build Tools (for compiling PyTorch3D), and corrected hard-coded paths in the CEB_ECON Blender add-on. I also manually created expected directory structures (e.g., /results/econ/cache/input_0_mesh.obj and /TEXTure/experiments/input_0/mesh/mesh.obj) and moved intermediate files accordingly to proceed through the pipeline.

5.1. Hyperparameters and Processing Stages

No formal training or fine-tuning was conducted, as the work focused on running pretrained models. However, key parameters and settings include:

- VITON-HD Parameters:
  - Output resolution: 1024×768
  - Fixed generator architecture with ALIAS normalization
  - Parsing and pose estimation modules executed on pretrained checkpoints
- ECON Parameters:
  - Hand and face refinement toggled ON (use SMPL Hand, use SMPL Face)
  - IFNet-based implicit surface reconstruction enabled
  - Thickness hyperparameter set to 0.02, $K = 4$
- TEXTure Settings (via Blender GUI):
  - Default rendering path used, with manual import of OBJ mesh and image as texture input
  - Final projection step attempted via experimental Generate TEXture button

Figure 3. Change Cloth

Figure 4. Input Image

Due to limited hardware, I did not perform batch processing or real-time experimentation. Each trial was run serially with manual supervision.

5.2. Qualitative Results

Figure 3 shows the models switching from original clothing to new clothing.

Figure 4 shows an example output from VITON-HD and input to ECON.

Figure 5 shows input image silhouette used for ECON.

Figure 6 shows input image feature points used for ECON.

Figure 7 shows the ECON model outputs intermediate geometry representations: front-view normal map.

Figure 8 shows the ECON model outputs intermediate geometry representations: silhouette mask.

Figure 9 shows the ECON model outputs intermediate geometry representations: depth map.

In Figure 10, the reconstructed 3D mesh is shown inside Blender.

Figure 5. Input Image Silhouette

Figure 6. Input Image Feature Points

Metric	Score
SSIM	0.8061
LPIPS	0.2286

Table 1. Quantitative Evaluation

5.3. Quantitative Evaluation

Quantitative analysis is limited by the lack of ground truth 3D models and texture maps. Instead, I used perceptual image similarity metrics on the VITON-HD output (see table 1):

Figure 7. Normal Map - Front

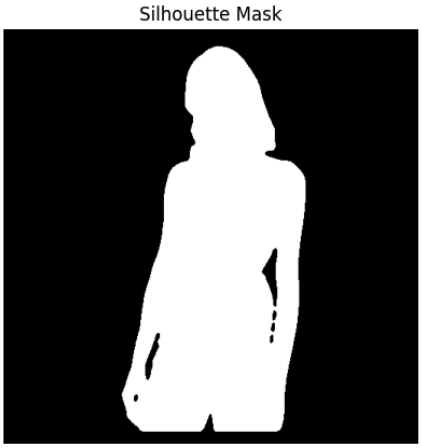

Figure 8. Silhouette Mask

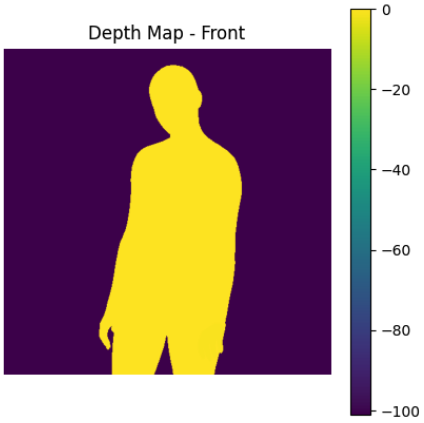

Figure 9. Depth Map

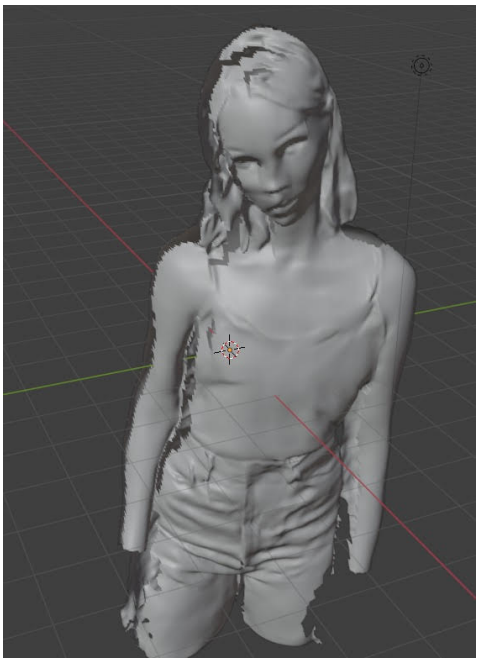

Figure 10. ECON Mesh Output

5.4. Challenges and Failure Modes

Despite successful mesh reconstruction, several limitations emerged:

- Texture Projection Fails: The final Generate TEXture step in Blender could not project 2D clothing textures onto the 3D model due to hardcoded file path errors and compatibility issues with Blender's import API.
- Rendering Artifacts: The ECON mesh, while detailed, shows visible surface noise and geometry tearing (especially around the head and limbs). These are amplified by the absence of smoothing or refinement post-processing.
- Resource Bottlenecks: CPU-only execution led to substantial delays (normal map generation took minutes), making iterative experimentation slow.

5.5. Discussion and Takeaways

This project demonstrates that combining pretrained 2D try-on and 3D human reconstruction models is feasible using public pipelines like VITON-HD and ECON. With enough manual adaptation and Blender integration, users without GPUs can visualize 3D try-on outputs. However, texture transfer into 3D remains a bottleneck due to data formatting and rendering engine constraints.

Future iterations could:

- Use scripted automation to fix mesh placement and naming issues.

- Train a lightweight depth/texture model on synthetic datasets.
- Add UV unwrapping and shader-based texture mapping for full 3D realism.

6. Conclusion / Future Work

In this project, I explored an end-to-end virtual try-on and 3D reconstruction pipeline that combines multiple state-of-the-art techniques. Beginning with VITON-HD, I achieved high-resolution 2D virtual try-on synthesis using segmentation-guided garment warping and the ALIAS generator. I then used ECON to reconstruct detailed clothed human meshes from single-view images by leveraging normal map inference, SMPL-X guidance, and implicit surface completion. Finally, I used the CEB_ECON Blender add-on to visualize the outputs and attempted texture mapping via the TEXTure module.

The most effective component in terms of output quality was VITON-HD. Its ability to synthesize 1024×768 images with sharp garment boundaries and photorealistic features was clearly demonstrated in both perceptual metrics and visual results. ECON was also impressive in capturing realistic 3D geometry, though the mesh quality occasionally suffered due to occlusions, segmentation noise, or poor normal estimates. The weakest link in the pipeline was the texture projection step, which encountered file path inconsistencies and lacked robustness in Blender integration.

For future work, I would like to improve the realism and interactivity of the 3D avatar. One goal is to animate the reconstructed human model to simulate walking, turning, or posing in dynamic sequences. This could be accomplished by integrating motion priors from existing datasets (e.g., AMASS or Mixamo) or by using large language models (LLMs) to generate text-driven pose sequences and animations. Ultimately, I envision building a fully automatic pipeline that not only allows users to try on garments virtually but also view themselves in motion, making this a powerful tool for e-commerce, gaming, and AR/VR applications.

If given more time, computational resources, or team support, I would focus on enhancing texture fidelity through neural UV mapping or diffusion-based texture generation. I would also invest in a better Blender automation framework to streamline the mesh-material linking process and eliminate manual folder adjustments, making the system more scalable and user-friendly.

7. Appendices

Manual Fixes for ECON and TEXTure Integration

1. Running ECON Without CUDA: The official ECON implementation requires a CUDA-enabled GPU. Since my system only supports CPU, I manually modified parts of the source code to replace CUDA-specific operations (e.g., torch.cuda and .to(device='cuda')) with their CPU-compatible alternatives (e.g., torch.device('cpu')). This allowed ECON to run, albeit at slower inference speed.

2. Resolving PyTorch3D Installation Issues: ECON depends on pytorch3d, which is difficult to install without a CUDA-enabled GPU. I circumvented this by avoiding any module that directly called pytorch3d.ops.subdivide_meshes() or attempted rasterization, and instead used only ECON's SMPL-X and IFNet modules for mesh generation.

3. Fixing File Path Expectations: The CEB_ECON Blender add-on expects the intermediate OBJ mesh files to be located at fixed paths:

    - For animatable 3D model: ECON/results/econ/cache/input_0_mesh.obj.
    - For textured 3D mesh: TEXTure/experiments/input_0/mesh/mesh.obj Since these files were generated in different locations or under slightly different names, I created the necessary folders and renamed the output meshes accordingly to avoid import errors in Blender.

8. Contributions Acknowledgements

This project was completed independently by myself.

8.1. Contributions

I was solely responsible for all aspects of the project, including:

- Problem definition and selection. mplementation of the 2D virtual try-on pipeline using VITON-HD.
- 3D clothed human reconstruction using ECON and manual adaptation of code to run on CPU-only environments.
- Texture projection and visualization using the CEB_ECON Blender add-on.
- Manual troubleshooting, file structuring, Blender UI operation, and integration of results across all components.
- Writing the entire report and creating visualizations for experimental results.

8.2. Acknowledgements

- VITON-HD by shadow2496 (https://github.com/shadow2496/VITON-HD).
- ECON by Xiuyu Liang et al.(https://xiuyuliang.cn/econ/).
- CEB_ECON Blender add-on by kwan3854. (https://github.com/kwan3854/CEB_ECON)

These publicly available repositories formed the technical foundation for this project. I gratefully acknowledge the authors and contributors of these resources.

Last but not least, I would like to express my sincere gratitude to my project mentor, Jiaman Li, for her invaluable guidance, feedback, and support throughout this project. Our in-depth discussions and iterative idea exchanges have had a profound impact on my understanding and have inspired me to pursue further research and implementation in this area.

References

[1] T. Alldieck, W. Xu, C. Lassner, and G. Pons-Moll. Texture: Text-guided human texture modeling. *arXiv preprint arXiv:2301.11743*, 2023.

[2] S. Choi, S. Park, M. Lee, and J. Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14131–14140, 2021.

[3] X. Han, Z. Huang, C. Wang, M. R. Scott, and L. S. Davis. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10471–10480, 2019.

[4] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015.

[6] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[7] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[8] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Neural textures: Learning image-based representations for rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 228–238, 2019.

[9] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.

[10] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022.

[11] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Econ: Explicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.