Run Time Optimization of SMPLX for Avatar Generation

Kaustav Mukherjee (kaustavmu) Adithya Narayan (anaraya2) Shaurye Aggarwal (shauryea)

Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA

Abstract

Recently, a plethora of pipelines have emerged to generate 3D clothed human avatars from single, in-the-wild images. However, all of them are limited to full-body, front-facing human images with minimal occlusions, objects, and simple poses. To address these limitations, we propose a two-part, inpainting and body fitting pipeline that addresses these issues. The inpainting pipeline uses keypoint detection and a novel keypoint estimation technique, uses LaMa for occluding object removal, Stable Diffusion with ControlNets for generation of missing areas, and a GAN inversion step to create a seamless, plausible human reconstruction. The body fitting pipeline uses an improved regressor and adds more losses to the iterative fitting stage to achieve a better human mesh fit in dynamic poses. Through qualitative comparisons, our pipeline shows improvements in mesh textures and SMPL-X fit over previous methods.

1. Introduction

There is an increasing demand for 3D clothed human models in a variety of applications, such as virtual and augmented reality (VR/AR), 3D printing, scene assembly, filmmaking, and video games. Since the manual process of creating such models requires considerable time, manpower, and specialized equipment, methods have arisen over time to expedite the process. Traditionally, such 3D models could only be created using multiview image inputs of clothed individuals [1]. However, such data is difficult and time-consuming to capture, and even harder to find in the wild. Therefore, there have been an increasing number of pipelines that attempt to create a clothed 3D human avatar from a single image [4, 11, 18, 25, 33, 34]. These can be created with in-the-wild images, removing the need for specialized data collection and streamlining the avatar generation process.

However, a majority of these methods are unable to account for certain in-the-wild scenarios that may occur. These include human-object interactions, occlusions, and highly dynamic poses, all of which create significant errors when utilizing these models. These are derived from all the different steps in the typical avatar generation procedure. First, there is a generative component that attempts to generate back or side views [11, 18, 33], during which interacted objects and occlusions tend to cause generation problems, and dynamic poses are not properly accounted for. Second, this is followed by iteratively fitting a base human 3D mesh [15], the most common of which is an SMPL [2, 19], during which errors for occluded and dynamic poses accumulate. Finally, there are a series of iterative normal map and texture estimation steps that create the final 3D clothed human avatar. In this step, object-interactions can result in odd model behaviors, and dynamic poses can lead to poor hidden-view normal estimation. All these effects are clear in Fig. 1, where it can be seen that smaller objects such as the soccer ball are absorbed into the human model and larger objects such as the man holding the box lead to extremely incorrect normal maps. Furthermore, in Fig. 2, it is also clear that dynamic poses lead to poor SMPL fits. To ameliorate this issue, we introduce additional processing steps to an existing pipeline to broaden the use-case of generating 3D avatars from in-the-wild images. Taking SIFU by Zhang et al. [33], which is able to create 3D avatars with realistic textures, capable of rigging and animation,

Figure 1. SIFU mesh reconstruction failures.

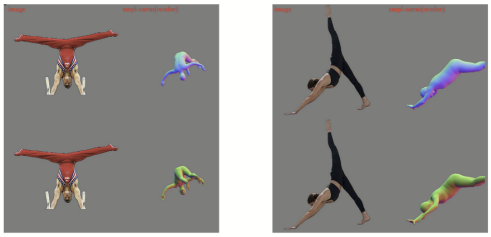


Figure 2. Pixie bodyfits struggle with dynamic poses.

3D printing, and complex 3D scene creation, we append two features: the first, an inpainting pipeline that uses diffusion [24, 26] and GAN models [9] to remove interacted objects and reconstruct occluded areas, and the second, a new regressor and iterative fitting step for the initial human 3D mesh fitting using PyMAF-x [30], along with additional depth and keypoint data to improve the iterative fitting loop. By creating these two solutions that are easily integrable into SIFU, our pipeline becomes more flexible and can be easily integrated into other, future human avatar generation methods with little extra work.

Therefore, we present Run Time Optimization of SMPLX for Avatar Generation, a novel pipeline that aims to extend the use-case of human avatar generation to more in-the-wild settings. While quantitative evaluation methods exist for 3D avatar generation, since existing datasets [6, 28] do not have meshes for in-the-wild images, we instead provide a qualitative evaluation that shows substantial improvements for SMPL prediction and normal map generation for human-object interactions, occluded images, and highly dynamic poses. Our key contributions include

- A novel keypoint estimation process using OpenPose [3] and XGBoost [5] to regenerate occluded humans.
- An inpainting pipeline to preprocess in-the-wild human images, suitable for any 3D avatar generation pipeline.
- An improved 3D human mesh fitting pipeline using the PyMAF-x regressor [30] and additional keypoint and depth constraints for the iterative fitting loop.

2. Related Work

Single-Image 3D Avatar Generation A majority of recent work done on 3D avatar generation concerns the use of single images to create 3D clothed human avatars [4, 11, 18, 25, 33, 34]. Early works, such as the seminal work PIFu by Saito et al. [25] use implicit functions to predict 3D geometry, but newer works [11, 18, 33] use explicit methods such as parametric body models [2] and fit them to the input images. These newer works also all have hidden-view generation steps that range from video diffusion models in HumanVDM [18], to control-net aided 2D diffusion in SiTH [11, 31], to transformer-based prediction in SIFU [33]. Where they also differ is their normal and RGB map prediction. While SiTH, SIFU, and Ultraman [4] all use neural networks to query points and determine normals and RGB values, HumanVDM uses 3D Gaussian Splatting [12] to create a riggable 3D avatar. Pipelines like SIFU also have additional texture refinement steps using text-based diffusion models to improve texture quality in hidden views. However, all of these methods only look at a narrow use-case of full-body clothed human images with little to no object interactions, and relatively simple poses found in datasets such as THuman and DNA-Rendering [6, 28], resulting in pipelines that malfunction for real in-the-wild images with human-object interactions, occlusions, and dynamic poses.

3D Human Mesh and Fitting Focusing in on explicit methods for 3D avatar generation, all of them have a requirement for parametric body models that can be fit to the input image. The first of these was a skinned, multi-person linear model, SMPL, by Loper et al. [19], soon after which came SMPLify [2], a CNN-based method to predict an SMPL fit for an input image. New models and methods have since released hand-in-hand, with SMPL-X [21] expanding on the SMPL model with additional parameters for facial and hand expression, and a new method SMPLify-X to account for and fit these features. Since then, new regressors have attempted to improve the SMPL-X fit, such as PIXIE [8], which is used by SIFU, which fits the hands, body, and face with sub-networks and combines the predictions for an improved fit. PyMAF and PyMAF-X [29, 30] use spatial feature pyramids and an additional mesh alignment iterative loop, while HybrIK and HybrIK-X [16, 17] utilize inverse kinematics to convert precise 3D keypoints to parametric human body meshes. Furthermore, newer and more complete pipelines like KBody [35] add further components to the iterative fitting loop to refine the body mesh after initial regression, such as depth, keypoint, camera, and silhouette losses. These techniques can be combined and improved upon to allow for a better parametric model fit in our proposed pipeline.

Human Image Generation and Inpainting Pre-existing works on human image inpainting include EXE-GAN [20], which is only trained on inpainting faces, and another method by Grigorev et al. [10] which aims to generate an image in a forward-facing pose from an incomplete image from a different pose. Neither of these are particularly relevant or well suited to the task at hand. A recent work, KBody [35], explores inpainting for the completion of human images before use in an SMPL-fitting pipeline, which is extremely relevant. It proposes the use of a pre-trained encoder [27] to project the input image into the StyleGAN-Human [9] latent space, and using a test-time finetuning technique known as pivotal tuning inversion (PTI) [23] to generate a complete image. However, with

testing, this process is unable to account for larger occlusions, dynamic poses, and some object interactions. Apart from GAN-based generation and inpainting, another possible avenue is the use of diffusion models [24] to generate missing areas. The first step in these pipelines is to utilise some segmentation model, such as YOLOv11 [13] or Segment Anything (SAM) [14] to identify missing areas, and then using both image and text conditioned diffusion to inpaint them. Further methods to control diffusion include controlnets [31], which propose the addition of one or more condition images to condition the diffusion process using an encoder and zero-convolutions during decoding steps. Segmentation masks, as discussed earlier, can be used as conditions, as well as human pose keypoints, such as those generated by OpenPose [3]. Other forms of inpainting include LaMa [26], which uses fourier transforms to generate missing areas of an image. While any single method here is insufficient to address the myriad of requirements for the inpainting pipeline, a combination of these can result in a more holistic solution.

3. Methodology

The overall pipeline is split into inpainting and SMPL-X estimation.

3.1. Inpainting

The full inpainting pipeline is shown in Fig. 3, and consists of four key different steps. First, a preprocessing step for segmentation and pose detection, then aligning the image and estimating missing keypoints, follow by mask- and pose-conditioned inpainting, and finally GAN-based inversion to smoothen the final image.

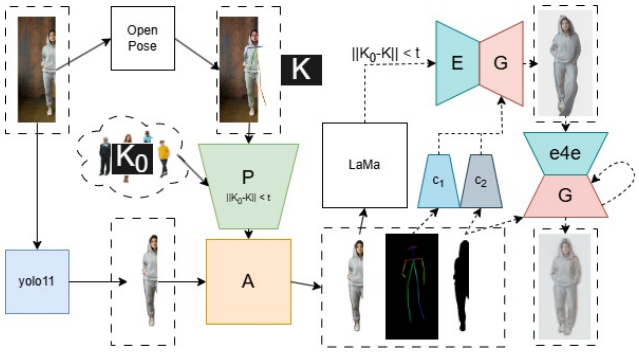


Figure 3. Full inpainting pipeline.

3.1.1 Preprocessing

To simplify and automate the preprocessing, off-the-shelf models are used for both segmentation and keypoint estimation. YOLOv11 [13] is pretrained for human bounding box and segmentation mask detection, as well as for a variety of object classes. This allows the model to identify any objects that the human in the image is interacting with, and generate a mask that can be used to remove said object in further steps. The keypoint detection model, OpenPose [3], is used as an an 18-point keypoint detector for face, hands, and body, and can determine keypoints for partial or occluded images, with the missing keypoints being generated by a later step.

3.1.2 Image Alignment and Keypoint Estimation

In order to input an image into StyleGAN-Human in the fourth step, the input image resolution must be 512×1024. Since the input to the pipeline may be incomplete or occluded, the image must first be properly resized, then aligned, and missing keypoints must then be generated. This is done using samples from StyleGAN-Human [27] and detecting keypoints to create a ground-truth dataset of keypoints, and also determine values for average locations of different keypoints, K_0. This is used to:

- Properly align the images to the output canvas by generating an affine transformation A.
- Since the StyleGAN-Human dataset only accounts for humans in forward-facing, non-dynamic poses, any input images in this pipeline will not be well-served by the keypoint generation and GAN inversion steps. Therefore, an L2 Norm is taken of the input image keypoints K and K_0, and a threshold t is used to determine whether a specific pose is too dynamic and will be exlucded from the aforementioned steps.
- For images within the threshold, train an XGBoost [5] model with K_0 to learn relationships between the keypoints. The keypoint generation process is treated as data imputation [7], meaning that missing keypoints are treated as missing data in an input dataset, and XGBoost is able to predict the remaining keypoints K_{in} using the trained model from K_0.

This process then outputs an aligned human image I_0, human mask S_H, object mask S_O, and keypoint image K_{in}.

3.1.3 Inpainting

The next key step is inpainting missing areas to remove occluding objects and regenerate limbs. Instead of a single inpainting step, a multi-step, two-model approach is used to achieve better results by capitalizing on the strengths of each model.

LaMa [26] is first used due to its ability to use contextual data around the object removal mask S_O to properly fill the missing area in the image I_0. In addition, LaMa comes with a refinement step that uses structural data from low-resolution generations to improve the high-resolution final image generation. This works particularly well for the

removal of occluding objects and generates image I_1, ready for the next inpainting step.

This is followed by the use of Stable Diffusion [24] with two controlnets [31] simultaneously - keypoint image K_in, and human segmentation mask S_H. This outputs an image I_2 that has a plausible human shape, but often contains seams between the original and generated area, extra limbs or fingers, or other odd generations that will affect the body fit and normal map generation in later steps. Therefore, GAN inversion is used to create a smoother final image.

3.1.4 GAN Inversion

The GAN inversion step is inspired by KBody [35], which directly uses GAN inversion on an input image. Here, it is used as the final step after larger occlusions, objects, and dynamic poses have been removed by the pipeline. The image I_2 is first passed through a pretrained e4e encoder [27] to map onto the latent space W of StyleGAN-Human [9], resulting in a latent representation w_2. After this, a final image I_3 can be generated, using generator weights θ, as $I_3 = G(w_2; \theta)$. However, instead of generating the final image by just using the default generator G, PTI [23] is used to fine-tune the generator at test time, which allows the regenerated image to better resemble the input image while having complete, coherent features. However, while PTI was designed for complete image editing, since the desired functionality here is to create a plausible inpainting result, the loss function is changed to reflect that. With tuned generator weights θ^*, where $I_4 = G(w_2; \theta^*)$. The original loss function of PTI is as follows:

$$\mathcal{L}_{pt} = \mathcal{L}_{LPIPS}(I_3, I_4) + \lambda_{L2}\mathcal{L}_{L2}(I_3, I_4) \quad (1)$$

It uses the L2 loss, as well as the perceptual similarity loss LPIPS [32]. For inpainting, a masked loss is used using S_H to calculate loss only for the area of the original input image, represented by $I_3 \odot S_H$, allowing the generator to create a plausible representation of the remaining image. While the latent space W_0 and representation w_0 remain the same, the generator tuning creates the final image I_F, which can then be used for the generation of human avatars.

3.2. SMPL-X Estimation

The SMPL-X estimation was enhanced by focusing on improvements in both the SMPL regressor and the iterative refinement process.

3.2.1 SMPL-X Regressor

To improve the performance of the SMPL-X body-fit, we replace the PIXIE[8] regressor with PyMAF-X[30]. Prior to this, we experiemnted with several regressors including Hybrik[16], and K-Body[35]. However both approaches

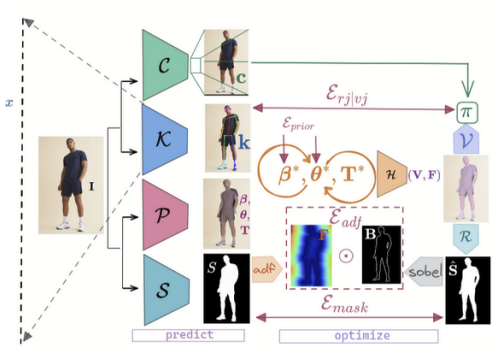


Figure 4. We use the system of losses recommended in K-Body [35].

were limited in specific ways. For instance, Before settling on PyMAF-X, we experimented with several other state-of-the-art regressors, including Hybrik[16], and K-Body[35]. While each of these methods demonstrated strengths in specific scenarios, they exhibited certain limitations. For instance, Hybrik, despite its kinematic modeling capabilities, failed to consistently capture fine-grained surface details. Similarly, K-Body performed well for rigid body postures but faced challenges with complex poses and self-occlusions.

PyMAF-X stood out as it leverages a hybrid representation that integrates parametric and non-parametric modeling to accurately capture detailed surface deformations while ensuring global consistency of the SMPL-X mesh. Unlike the other methods, PyMAF-X employs a multi-level feature aggregation approach, iteratively refining body shape and pose estimation. This approach significantly enhances robustness and accuracy, particularly in challenging scenarios such as extreme poses, occluded body parts, or noisy inputs.

Furthermore, PyMAF-X's hierarchical optimization pipeline, which combines pixel-aligned features with global shape priors, ensures anatomically plausible SMPL-X reconstructions. These improvements made PyMAF-X the most reliable choice for our application, providing an accurate and robust foundation for downstream tasks such as pose transfer and motion retargeting.

3.2.2 SMPL-X Iterative Fitting

We use a variation of the system of losses shown in Fig. 4, incorporating a mixture of losses alongside the SMPL loss and the silhouette loss used in the PyMAF-X framework. Much like most SMPL works [15][30][19], we use L_{SMPL} and $L_{silhouette}$ components to refine the body pose and shape parameters. However, inspired by K-Body[35] we also include an edge loss (L_{Sobel}) and a distance map based loss ($L_{shillouette}$). The overall loss function $\mathcal{L}_{\text{total}}$ is

(a) Inpainting pipeline image results.

(b) Mesh reconstruction with inpainting.

Figure 5. Inpainting and mesh reconstruction results.

formulated as a weighted combination of these losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{SMPL}}\mathcal{L}_{\text{SMPL}} + \lambda_{\text{silhouette}}\mathcal{L}_{\text{silhouette}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{Sobel}}\mathcal{L}_{\text{Sobel}}.$$

The distance-map loss $\mathcal{L}_{\text{distance}}$ helps refine the spatial alignment of body parts. It computes the difference between predicted and ground truth distance maps D_{pred} and D_{GT}, ensuring accurate body surface reconstruction:

$$\mathcal{L}_{\text{distance}} = \frac{1}{P}\sum_{p=1}^{P} \|D_{\text{pred}}(p) - D_{\text{GT}}(p)\|_2^2$$

Finally, the Sobel loss $\mathcal{L}_{\text{Sobel}}$ leverages edge information by penalizing differences in gradients between the predicted and ground truth silhouettes. This helps refine edge regions, improving alignment at boundaries:

$$\mathcal{L}_{\text{Sobel}} = \|\nabla_x S_{\text{pred}} - \nabla_x S_{\text{GT}}\|_2^2 + \|\nabla_y S_{\text{pred}} - \nabla_y S_{\text{GT}}\|_2^2$$

where ∇_x and ∇_y are horizontal and vertical gradients.

We observed that while the camera-based loss and keypoint loss did not provide significant improvements over the baseline, the inclusion of distance-map and Sobel losses notably enhanced the body parameter fit. This combination of losses helps achieve more accurate body pose estimation and better alignment with the target silhouette, resulting in improved performance over the baseline. Furthermore, since we observed that the magnitude L_{Sobel} and L_{dist} was one order of 10 larger than the other loss components. To counteract this, $\lambda_{dist} = 0.01$,$\lambda_{Sobel} = 0.1$, $\lambda_{silhouette} = 1$ and $\lambda_{SMPL} = 1$ were selected.

4. Results

To assess our method, we implement our inpainting and body fitting pipelines within SIFU [33] and perform qualitative evaluations on the SMPL fit and normal maps on a variety of in-the-wild images found online. This is in lieu of a quantitative evaluation due to the lack of human-to-mesh datasets with objects, occlusions, and highly dynamic poses. Through these qualitative evaluations, we demonstrate the effectiveness of our method in dealing with human-object interactions, occlusions, and highly dynamic poses compared to SIFU.

4.1. Implementation

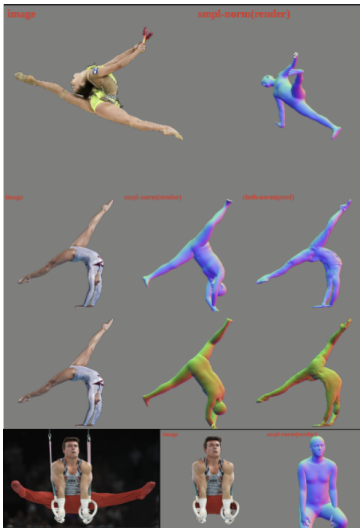
The inpainting pipeline is appended in front of the SIFU pipeline, such that outputs of this pipeline are input to SIFU. XGBoost 2.1.3 is used for keypoint estimation. LaMa is run with the big-lama checkpoint, with refinement set to True and refinement iterations set to 50. ControlNet is implemented in Stable Diffusion 1.5 using HuggingFace, which is provided directly by the authors of the paper. Finally, the GAN inversion pipeline is adapted from code in the StyleGAN-Human repository. Due to computational constraints, the full inpainting pipeline was run on CPU, with an average runtime of 30 minutes per image. PyMAF-X is implemented in place of PIXIE, and additional losses are added to the iterative fitting of SIFU, and run for 100 iterations, running in under a minute per image using an NVIDIA 3090 GPU.

4.2. Inpainting Pipeline

Three categories of images were tested for inpainting: small hand-object interactions, images with cut-offs, and large occluding objects. From Fig. 5a, it can be seen that for small hand-object interactions, convincing outputs with few artifacts can be achieved by the pipeline. As these images have most of the keypoints and fairly dynamic poses, the keypoint estimation and GAN inversion steps are automatically skipped, resulting in a very realistic output due to the lower amount of processing needed. Images with missing body sections go through the keypoint estimation and GAN

(a) Our implementation vs. SIFU.

(b) More fits on dynamic poses.



(c) Various failure cases.

Figure 6. PyMAF-X body fit regression results.

inversion steps, resulting in a final product with more artifacts. The results from the diffusion step have odd generations, especially near the legs, and while the GAN inversion step produces a more anatomically realistic final result, the pictures are blurry. Especially for the image where the right side is occluded, significant facial detail is lost. Finally, images with larger occlusions have significantly worse performance. It can be observed poor diffusion outputs lead to significantly noisier final products after GAN inversion.

As the goal of this pipeline is to produce an improved normal map for 3D clothed avatars, testing the inpainted images with the SIFU pipeline shows its efficacy in removing errors. In Fig. 5b, SIFU is unable to generate correct normal maps for occluded and incomplete images, resulting in odd and incomplete textures for not only the affected area but also visible areas such as the neck and face. Our pipeline allows for the generation of a more realistic, complete, and error-free normal map. However, there are still issues with backgrounds being absorbed into the body, as well as lower detail in the face and clothing.

4.3. SMPL-X Fitting

The use of PyMAF-X shows clear qualitative improvements over the baseline on highly dynamic poses. Fig. 6a shows significantly better fits for limbs, along with smaller details like head alignment. More examples are seen in Fig. 6b, where various dynamic poses in athletic environments are properly fit to by our implementation. However, certain failure cases are visible in Fig. 6c. Very flexible poses are still problematic, along with dynamic poses where the limbs can be easily swapped. There are also issues with background removal leading to poor fits.

5. Conclusion

In this work, we attempt to bring single-image to 3D clothed avatar reconstruction to more in-the-wild scenarios by accounting for human-object interaction, large occlusions, and dynamic poses. We leverage off-the-shelf segmentation and pose detection models, and create an inpainting pipeline with keypoint estimation, diffusion-based object removal and inpainting, and GAN inversion to create a plausible human image based on the occluded input. We also implement PyMAF-X [30] to replace the body fit regressor, improving the fit for dynamic poses. We then implement both changes in the pre-existing SIFU [33] to generate improved SMPL fits and normal maps. We qualitatively evaluate these methods against SIFU to show its success in accounting for the aforementioned in-the-wild scenarios when reconstructing 3D human avatars from single images.

6. Future Work

Due to the lack of previous work tackling these exact problems, much work can be done to improve these results. Fine-tuning StyleGAN-Human [9] on dynamic poses will allow for reconstructions on more poses. Using ensembles of face and hand keypoint detection, prediction, and generation models can improve details in those areas. Failure cases for body fitting dynamic poses from certain angles can be addresed by a more robust regressor and additional losses in the iterative fitting process. Furthermore, better background separation will prevent cases of limbs being removed when removing backgrounds in dynamic poses. Finally, we could explore performance on standard datasets like CAPE [22], and also explore using synthetic data to better evaluate this model on harder poses.

References

[1] Naveed Ahmed, Edilson de Aguiar, Christian Theobalt, Marcus Magnor, and Hans-Peter Seidel. Automatic generation of personalized human avatars from multi-view video. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, page 257–260, New York, NY, USA, 2005. Association for Computing Machinery. 1

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, 2016. 1, 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2, 3

[4] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail, 2024. 1, 2

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM, 2016. 2, 3

[6] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering, 2023. 2

[7] Yongshi Deng and Thomas Lumley. Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, 33(2):352–363, 2024. 3

[8] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation, 2021. 2, 4

[9] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation, 2022. 2, 4, 6

[10] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided image generation, 2019. 2

[11] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion, 2024. 1, 2

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 2

[13] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. 3

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[15] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop, 2019. 1, 4

[16] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2, 4

[17] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 2

[18] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models, 2024. 1, 2

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael Black. Smpl: a skinned multi-person linear model. 2015. 1, 2, 4

[20] Wanglong Lu, Hanli Zhao, Xianta Jiang, Xiaogang Jin, Yongliang Yang, Min Wang, Jiankai Lyu, and Kaijie Shi. Do inpainting yourself: Generative facial inpainting guided by exemplars, 2022. 2

[21] 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, 2019. 2

[22] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 6

[23] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images, 2021. 2, 4

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2, 3, 4

[25] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization, 2019. 1, 2

[26] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. 2, 3

[27] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021. 2, 3, 4

[28] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2

[29] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop, 2021. 2

[30] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10), 2023. 2, 4, 6

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3, 4

[32] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 4

[33] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction, 2024. 1, 2, 5, 6

[34] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2020. 1, 2

[35] Nikolaos Zioulis and James F. O'Brien. Kbody: Towards general, robust, and aligned monocular whole-body estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 6215–6225. IEEE, 2023. 2, 4