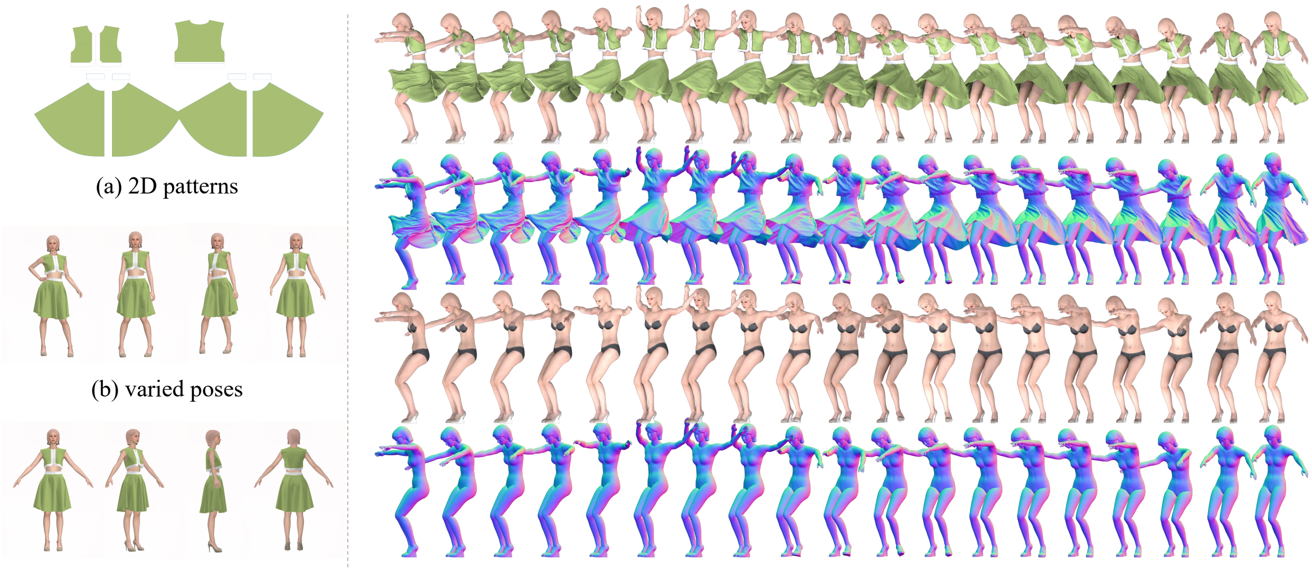
CLOTH4D: A Dataset for Clothed Human Reconstruction

Xingxing Zou[1,3*], Xintong Han[2*], Waikeung Wong[3,1†]
[1]Laboratory for Artificial Intelligence in Design, [2]Huya Inc.
[3]School of Fashion and Textiles, The Hong Kong Polytechnic University
aemika.zou@connect.polyu.hk, hanxintong@huya.com, calvin.wong@polyu.edu.hk


(a) 2D patterns

(b) varied poses

(c) multi views

(d) series of paired dynamic textured meshes

Figure 1. An instance in CLOTH4D. Given (a) hand-crafted 2D sewing patterns, clothes are simulated with virtual fashion software to have (b) varied poses, (c) multiple views, and (d) high-quality and physically plausible series of paired dynamic 3D meshes with textures.

Abstract

Clothed human reconstruction is the cornerstone for creating the virtual world. To a great extent, the quality of recovered avatars decides whether the Metaverse is a passing fad. In this work, we introduce CLOTH4D, a clothed human dataset containing 1,000 subjects with varied appearances, 1,000 3D outfits, and over 100,000 clothed meshes with paired unclothed humans, to fill the gap in large-scale and high-quality 4D clothing data. It enjoys appealing characteristics: 1) Accurate and detailed clothing textured meshes—all clothing items are manually created and then simulated in professional software, strictly following the general standard in fashion design. 2) Separated textured clothing and under-clothing body meshes, closer to the physical world than single-layer raw scans. 3) Clothed human motion sequences simulated given a set of 289 actions, covering fundamental and complicated dynamics. Upon CLOTH4D, we novelly designed a series of temporally-aware metrics to evaluate the temporal stability of the generated 3D human meshes, which has been overlooked previously. Moreover, by assessing and retraining current state-of-the-art clothed human reconstruction methods, we reveal insights, present improved performance, and propose potential future research directions, confirming our dataset's advancement. The dataset is available at [1].

1. Introduction

As we enter the volumetric and XR content era, researchers have been trailblazing their way into the Metaverse. With the converging of technologies and practical applications, *e.g*., fashion NFTs (non-fungible tokens), immersive AR and VR, and games, clothed human reconstruction demands are rapidly growing. While current research has made astonishing results in creating digital hu-

[1] www.github.com/AemikaChow/AiDLab-fAshIon-Data

*X. Zou and X. Han contribute equally. † Corresponding author.

Table 1. Comparisons of CLOTH4D with existing representative datasets. Gray color indicates synthetic datasets generated with graphics engines. #Subjects: number of peoples in different appearances; #Action: number of actions adopted; #Scans: numbers of 3D meshes; 2D Pattern: 2D clothing pattern; TexCloth: with textured clothed model; TexHuman: with textured naked human model. w/ SMPL: with registered SMPL [33] parameters. Public: publicly available and free of charge. Photorealistic: whether the images in the dataset are realistic. -: not applicable or reported. CLOTH4D presents more desirable characteristics compared with others.

Dataset	#Subjects	#Action	#Scan	2D Pattern	TexCloth	TexHuman	w/ SMPL	Public	Photorealistic
BUFF [52]	6	-	13.6k	-	✓	-	✓	✓	✓
RenderPeople [1]	-	-	825	-	✓	-	✓	-	✓
DeepWrinkles [30]	2	2	9.2k	-	✓	-	-	-	✓
CAPE [35]	15	600	140k	-	-	-	✓	✓	✓
THuman2.0 [51]	200	-	525	-	✓	-	✓	✓	✓
DRAPE [16]	7	23	24.5k	-	-	-	-	-	-
Wang et al. [47]	-	-	24k	✓	✓	-	✓	✓	-
3DPeople [40]	80	72	-	-	✓	-	-	-	✓
DCA [43]	-	56	7.1k	-	-	-	✓	-	-
GarNet [17]	600	-	18.8k	-	-	-	✓	✓	-
TailorNet [38]	9	-	5.5k	-	✓	-	✓	✓	-
Cloth3D [8]	8.5k	7.9k	2.1M	-	✓	-	✓	✓	-
Cloth3D++ [36]	9.7k	8k	2.2M	✓	✓	✓	✓	✓	-
CLOTH4D	1k	289	100k	✓	✓	✓	✓	✓	✓

mans, these reconstructed meshes have issues, *e.g.*, flexible body motions and diverse appearances, owing to the lack of datasets with richness in clothing and realistic dynamics of garments. To this end, we introduce CLOTH4D, an open-sourced dataset facilitating physically plausible dynamic clothed human reconstruction.

Prior to us, many datasets have been collected, and we sort out them in Table 1. Currently, scanned datasets are widely adopted as they are photorealistic and can be easily processed to watertight meshes, which does an excellent favor for current deep models to learn an implicit function (*e.g.*, signed distance function) followed by marching cubes [34] for surface reconstruction. However, it is born with some weaknesses: 1) Scanned meshes are single-layer and inherently fail to capture the space between clothing and skin surface. Thus, body shape under clothing cannot be accurately inferred, let alone the multi-layer and thin clothing structures as in the real physical world. 2) It is time-consuming and expensive to obtain high-quality and large-scale temporal scanned sequences (*i.e.*, 4D scanned sequences) due to the limited efficiency and precision of 4D scanners, especially for complicated clothing and large motions. Although synthetic datasets can to some extent overcome these limitations, existing synthetic datasets are either of small scale in terms of appearances and motions or are highly unrealistic. Moreover, many datasets are not made publicly available and free.

In contrast, CLOTH4D possesses several attractive attributes: 1) We made great efforts to the diversity and quality of clothing. All clothes are manually designed in CLO [3] and cater to the requirement of the fashion industry. 2) Meshes in CLOTH4D are clothing/humans separated. Such flexibility makes studying and modeling the relations and interactions between clothing simulation and body movement possible. 3) CLOTH4D provides plenty of temporal motion sequences with realistic clothing dynamics. As the human body moves, the dressed clothing, *e.g.*, the skirt in Figure 1, naturally deforms. 4) The dataset is large-scale and openly accessible.

To demonstrate the advantages of CLOTH4D, we use it to evaluate the state-of-the-art (SOTA) clothed human reconstruction methods. In addition to the generally adopted static evaluation metrics, we propose a set of temporally-aware metrics to assess the temporal coherence in a video inference scenario thanks to the rich and true-to-life 4D synthetic sequences in the dataset. Quantitative and qualitative results of SOTA methods on CLOTH4D suggest that our dataset is challenging and the temporal stability of the reconstructed mesh is vital for evaluating the perceptual quality. Meanwhile, we retrain SOTA methods on CLOTH4D, revealing interesting observations of how they perform on multi-layer meshes with thin clothing structures. With in-depth analysis and a summary of challenges for the existing approaches, CLOTH4D makes an essential step toward more realistic reconstructions of clothed humans and stimulates several exciting future work directions. All in all:

• We contribute CLOTH4D, a large-scale, high-quality, and open-accessible 4D synthetic dataset for clothed human reconstruction.

• We introduce a series of temporally-aware metrics to evaluate the reconstructed performance in the aspect of temporal consistency.

• With the proposed dataset and metrics, we thoroughly analyze the pros and cons of SOTAs, summarize the existing challenges toward more realistic 3D modeling, and propose potential new directions.

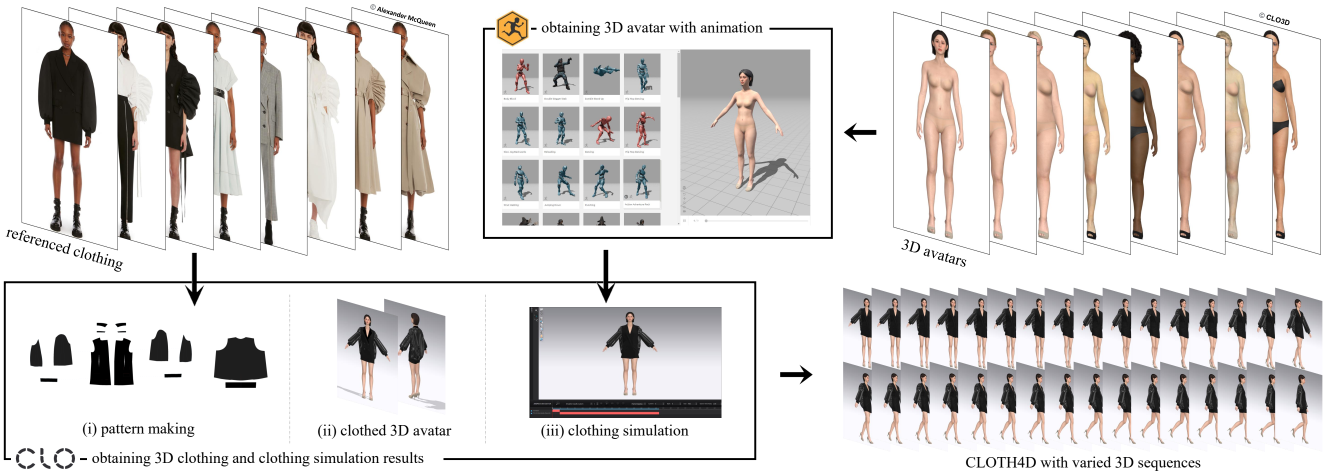

© Alexander McQueen
obtaining 3D avatar with animation
© CLO3D
referenced clothing
3D avatars
(i) pattern making
(ii) clothed 3D avatar
(iii) clothing simulation
CLO - obtaining 3D clothing and clothing simulation results
CLOTH4D with varied 3D sequences

Figure 2. Pipeline for creating instances in CLOTH4D, which primarily adopts CLO for clothing design and simulation, Mixamo for animation, and Blender for processing and exporting meshes.

2. Related Work

Clothed Human Reconstruction. Clothed human reconstruction aims to recover a 3D mesh from a monocular person image. Estimating dressed people by modeling clothing geometry as 3D displacement on top of a parametric body model (*e.g.*, SMPL [33]) is the leading solution to deal with this task [5, 6, 28, 31, 39, 48]. By transferring the skinning weights from the body model to the offset clothing mesh, the reconstructed clothed mesh can be readily deformed and animated in the same way as the under-clothing 3D parametric body model. However, it assumes the clothed human to have the same topology as the naked body, leading to unsatisfactory reconstructions of long hair, skirts, dresses, *etc*. Although methods such as [9, 22, 26, 38] try to isolate the reconstruction of clothing by constructing category-specific clothing templates or statistical models, they fail to generalize to unseen or complex types of clothing.

As another line of research, deep implicit function networks have drawn broader attention recently [7, 15, 20, 21, 24, 41, 42, 54]. PIFu [41] conditions on pixel-aligned features to build deep implicit functions for reconstructing human meshes, and PIFuHD [42] goes a further step towards enhancing 3D geometric details by predicting front and back normals. ARCH [24] and ARCH++ [21] enable animating the reconstructed meshes by deforming the semantic information into the SMPL canonical space. More recently, leveraging SMPL body models as prior, PaMIR [54] and ICON [49] further improve the reconstruction quality, especially on challenging poses. PHORHUM [7] achieves more accurate results via jointly estimating geometry, albedo, and shading information. It is worth noting that most works rely on training data that are not available free of charge, making an open-sourced dataset of clothed avatars of great significance to advance research in 3D body/clothing modeling.

Clothed Human Datasets. As summarized in Table 1, existing clothed human datasets can be divided into two types, *i.e.*, scanned datasets and synthetic datasets. The former [10, 25, 44, 46] utilizes multiple synchronized cameras to capture motions that have difficulties enriching the scalability and diversity and obtaining highly accurate 4D ground truth. Moreover, it inherently cannot mimic the layering of clothing. The synthetic datasets mitigate these limitations. However, existing synthetic datasets, no matter whether static [17, 38, 47] or dynamic [8, 16, 39, 43], either only contain a few clothing types or are highly unrealistic. Cloth3D++ [36] is developed from Cloth3D [8], which contains a total of 2.2 million scans, covering 9.7k subjects dressed in 12.9k clothing, which is the most advanced in scale. However, the clothing created based on garment templates are only base patterns with an immense gap compared to clothing in real life. To intuitively demonstrate the advantages of CLOTH4D, we put visual comparisons of these datasets in Figure A in the supplementary material.

3. CLOTH4D Dataset

We depict the pipeline of creating CLOTH4D in Figure 2, including (1) preparing the referenced clothing images and the unclothed 3D human avatars; (2) uploading 3D human avatars to Mixamo to obtain the FBX files with various animations; (3) designing 3D clothing in CLO, integrating the FBX files obtained in (2), and conducting clothing simulation; and (4) exporting the sequenced mesh files.

Clothes and Models. Clothes are manually created by professional fashion designers using CLO by producing the 2D garment patterns and then auto-simulating them to 3D. We put a video illustrating the production process of a 3D garment in the supplementary material (Video A) instead of demonstrating details in the paper. CLOTH4D covers 1,000 different 3D outfits spanning over 500 prints and 50 fabrics. Over 40% of the clothes are designed referring to the newest collections of varying design houses (*e.g.*, Prada, Moschino, and Alexander McQueen) to ensure visual realism, variance, and fashionability. In this paper, we focus on women's wear owing to its diversity covering most characteristics of garments. For avatars, CLO provides human avatars with varied physical appearances in terms of hairstyles, faces, skin colors, body figures, *etc*., and we directly adopt these available avatars.

Animations and Simulations. For each human avatar, we use Mixamo [2] for rigging and generating motion sequences. Given a motion sequence and a 3D garment, CLO runs a clothing simulation of the 3D garment according to the motion sequence at 30fps. A total of 289 animations are utilized, such as Belly Dance, Offensive Idle, Jumping Rope, *etc*. Unlike previous work [8] directly uses Blender for cloth simulation, CLO simulates clothing with richer details and dynamics. The clothes exhibit wrinkles and fold with the body's movement in a natural and physically plausible way, especially for skirts or dresses that differ a lot from the body topology as shown in Figure 1.

Paired Multi-Modal Data. Then, we can obtain paired data in the following representations: a 2D clothing pattern, 3D mesh sequences (clothed and naked human meshes with corresponding fitted SMPL parameters/meshes, and separate clothing meshes), and a UV texture map. The textured mesh can be rendered into multi-view normal images, depth images, and RGB images given varying light conditions. We also translate all these dynamic meshes to watertight with simplification using [23], thus they can be readily used to train an implicit function for mesh reconstruction. The number of triangles ranges from $170K$ to $4M$ for simulation and becomes $200K$ after simplification. Besides human reconstruction, many other tasks could also benefit from this paired data (*e.g*., clothing capture [50], human pose transfer [18], and fashion-related tasks [19, 56, 57]).

4. Evaluations

In this section, we evaluate the state-of-the-art clothed human reconstruction methods on CLOTH4D to demonstrate new insights that the dataset can provide. Further, we retrain SOTA methods on CLOTH4D and make several interesting observations of how they perform on multi-layer meshes with thin clothing structures. We also present the challenges for existing approaches and propose potential research directions with a comprehensive analysis.

4.1. Baselines

We mainly report results of four SOTA approaches, PIFu [41], PIFuHD [42], PaMIR [54], and ICON [49] owing to other works do not release their codes or models, such as PHORHUM [7], ARCH++ [21], or which have already been extensively compared with the listed methods above. We use PIFu, PIFuHD, PaMIR, and ICON to denote their released pretrained testing models. PaMIR^{gt} and ICON^{gt} indicate the testing results using the fitted ground truth SMPL mesh as conditions rather than the one predicted using off-the-shelf human mesh recovery (HMR) methods (GraphCMR [29] for PaMIR, and PyMAf [53] for ICON as in their released code). Based on the characteristics of these methods, they can be divided into three types: 1) pixel-aligned methods (PIFu, PIFuHD); 2) GT-SMPL-guided + pixel-aligned methods (PaMIR^{gt} and ICON^{gt}); 3) HMR-SMPL-guided + pixel-aligned methods (PaMIR and ICON).

Furthermore, we retrain PIFu, PaMIR, and ICON on CLOTH4D, which are denoted as PIFu$_{clo}$, PaMIR$_{clo}$, and ICON$_{clo}$, respectively. For these retrained models, we follow the re-implementation setting introduced in ICON [4], which allows us to train all these baselines with the same training protocol and hyper-parameters for a fair comparison. Similarly, PaMIR$^{gt}_{clo}$, and ICON$^{gt}_{clo}$ are tested with the ground truth SMPL fits. We also use the cloth-refinement module in the ICON's released code for post-processing.

4.2. Datasets and Metrics

Datasets and implementation details. We organize the sequences in CLOTH4D into a 80%/10%/10% train/val/test split. We render each mesh into 8 views using a weak perspective camera and pre-computed radiance transfer [45] with dynamic light conditions following [41, 49]. All rendered images are 512×512. The 2D keypoints used in all methods are generated by OpenPose [11]. In addition, we also evaluate all models on CAPE [35] test set adopted in [49] to investigate the generalization ability.

Static Metrics. We report the quantitative results on normal reprojection error, Chamfer distance, and P2S distance for evaluation as [15, 41, 42, 49]. As all compared methods use a weak perspective or orthographic camera, the estimated meshes may not be well aligned with the ground truth meshes in the z-direction (*i.e*., view direction). Thus, we shift the estimated meshes to have the same z-axis mean as the ground truths following [15] for a fair comparison.

Temporal Metrics. The aforementioned static metrics ignore the temporal consistency of the reconstructed meshes across time, which is essential for real-time applications since meshes presenting jitters and flickers highly affect the perceptual quality. Thanks to the rich temporal dynamics provided in CLOTH4D, we are the first to introduce temporally-aware metrics to evaluate the temporal coherence of the generated mesh sequences. Referring to temporal metrics SSDdt and dtSSD used in video matting tasks [14, 32], we compute two metrics measuring the temporal coherence of the predicted mesh normal:

$$\text{Normals}_{ddt} = \frac{1}{T}\sum_t \left| \left(\mathcal{N}_t^{pr} - \mathcal{N}_t^{gt}\right)^2 - \left(\mathcal{N}_{t+1}^{pr} - \mathcal{N}_{t+1}^{gt}\right)^2 \right|, \quad (1)$$

$$\text{Normals}_{dtd} = \frac{1}{T}\sum_t \left| \left(\mathcal{N}_t^{pr} - \mathcal{N}_{t+1}^{pr}\right)^2 - \left(\mathcal{N}_t^{gt} - \mathcal{N}_{t+1}^{gt}\right)^2 \right|, \quad (2)$$

where T is the length of the sequence. $\mathcal{N}_t^{pr}$ and $\mathcal{N}_t^{gt}$ denote the rendered normal images from the predicted mesh

Table 2. Quantitative evaluation on CLOTH4D. PaMIR^{gt} and ICON^{gt} denote that the fitted ground truth SMPL is used during the inference. Gray color indicates the results trained on CLOTH4D.

Method	PIFu	PIFuHD	PaMIR	PaMIR^{gt}	ICON	ICON^{gt}	PIFu$_{clo}$	PaMIR$_{clo}$	PaMIR$^{gt}_{clo}$	ICON$_{clo}$	ICON$^{gt}_{clo}$
Normals ↓	0.182	0.181	0.224	0.145	0.211	**0.118**	0.150	0.230	0.114	0.198	0.103
P2S ↓	3.911	3.518	3.827	2.641	4.660	**2.473**	2.793	5.037	2.619	3.711	2.068
Chamfer ↓	3.578	2.487	4.157	2.487	4.196	**1.631**	2.412	4.186	1.618	3.499	1.367
Normals$_{ddt}$ ↓	0.008	**0.007**	0.032	0.025	0.013	0.030	0.009	0.021	0.033	0.013	0.035
Normals$_{dtd}$ ↓	0.025	**0.023**	0.043	0.035	0.034	0.028	0.017	0.040	0.027	0.038	0.033
P2S$_{ddt}$ ↓	0.222	**0.167**	0.253	0.247	0.383	0.369	0.185	0.417	0.402	0.321	0.367
P2S$_{dtd}$ ↓	0.733	0.578	1.025	0.702	0.979	**0.554**	0.548	0.890	0.536	0.878	0.526
Chamfer$_{ddt}$ ↓	0.207	**0.181**	0.331	0.316	0.359	0.358	0.157	0.350	0.382	0.331	0.369
Chamfer$_{dtd}$ ↓	0.729	0.576	1.017	0.701	0.974	**0.553**	0.546	0.886	0.536	0.873	0.525

Table 3. Quantitative evaluation on CAPE. Table notations are the same as Table 2.

Method	PIFu	PIFuHD	PaMIR	PaMIR^{gt}	ICON	ICON^{gt}	PIFu$_{clo}$	PaMIR$_{clo}$	PaMIR$^{gt}_{clo}$	ICON$_{clo}$	ICON$^{gt}_{clo}$
Normals ↓	0.161	0.160	0.183	0.086	0.160	**0.056**	0.164	0.176	0.093	0.156	0.077
P2S ↓	4.259	3.795	3.840	1.193	4.014	**1.067**	4.652	4.235	1.491	3.304	1.193
Chamfer ↓	4.204	3.927	4.258	1.654	3.962	**1.038**	4.381	4.080	1.427	3.544	1.397

and the ground truth mesh at time step t, respectively. The subscript *ddt* is short for *distance delta time*, which captures the stability of errors between two consecutive meshes. And *dtd* (*delta time distance*) penalizes large temporal change of the prediction with respect to the change of the ground truth. These two metrics indicate unstable mesh variations and ignore temporally coherent errors [14]. The *ddt* and *dtd* of Chamfer and P2S distances are similarly defined. More details can be found in the supplementary material.

4.3. Baseline Evaluation

Quantitative results. Table 2 gives quantitative results on the CLOTH4D test set using the evaluation metrics described in Section 4.2. As indicated in the non-gray part, we made the following observations:

1) In terms of static metrics, ICON^{gt} > PaMIR^{gt} > PIFuHD > PIFu > PaMIR > ICON. *I.e.*, GT-SMPL-guided + pixel-aligned methods > pure pixel-aligned methods > HMR-SMPL-guided + pixel-aligned methods.

2) With the strong guidance of ground truth SMPL mesh, the performance of ICON^{gt} and PaMIR^{gt} significantly improves compared to their counterparts with estimated SMPL (*i.e.*, ICON and PaMIR). However, ground truth SMPL meshes are unavailable at test time, which suggests that previous comparisons [15,49] between GT-SMPL-based methods and others may be unfair. And pure pixel-aligned methods may be even more favorable than SOTA SMPL-based methods for the in-the-wild scenario.

3) From the perspective of temporally-aware metrics, pure pixel-aligned methods have higher reconstruction stability (see Figure 4). We attribute this to the fact that ICON and PaMIR strongly rely on the SMPL body prior, thus failing to generate far-from-the-body clothes (*e.g.*, skirts and dresses) that present rich temporal dynamics as shown in Figure 1. Plus, the jittery and unstable pose estimation of the off-the-shelf HMR methods further prevents ICON and PaMIR from generating temporally coherent results.

4) ICON outperforms PaMIR as ICON better models mesh-based local features while PaMIR depends more on global information. Moreover, PaMIR loses high-frequency details due to the limited resolution of its volumetric representation. Our observations on CAPE are in line with previous works. For simplicity, we do not expand the narrative.

Qualitative results. We present the qualitative results on CLOTH4D in Figure 3 and draw the following insights that have not been fully explored since there are no such large-scale and diverse datasets like CLOTH4D.

1) Global shape *vs*. local details. All baselines can reconstruct the overall shape conditioned on the input RGB image. PIFuHD presents the finest details, followed by ICON, PaMIR, and PIFu, as PIFuHD enlarges the spatial resolution of pixel-aligned features and ICON takes advantage of mesh-based local features (signed distance, surface normal, *etc*.). However, focusing on local features suffers from overfitting and poor generalization to complicated clothing (*e.g.*, incomplete dress in the 3rd and 5th examples in Figure 3) and large motions (*e.g.*, artifacts in the arm regions in the 2nd, 3rd, 6th, and 8th examples in Figure 3). Comparatively, thanks to the global feature encoder, PaMIR, and PIFu can generate more holistic clothing but sacrifice details. Thus, it is an important future research direction to *explore better strategies for balancing the local and global reconstruction quality*.

2) Human body priors. The last three rows in Figure 3 and Figure 4 (also shows side views) present results on relatively challenging poses. ICON^{gt} and ICON robustly recover the poses, while PIFu, PIFuHD, and PaMIR are prone to producing broken limbs or anatomically improbable shapes to different extents. For the side view, we can find that PaMIR^{gt} and ICON^{gt} are more similar to the

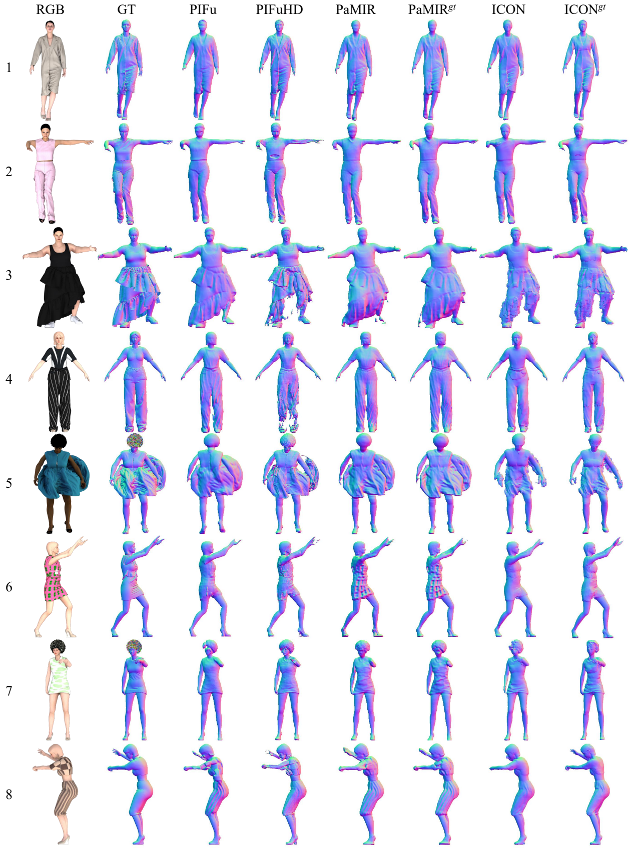


Figure 3. Qualitative results on CLOTH4D.

ground truth mesh, which is unsurprising as they are given the ground truth SMPL as prior. Comparatively, PaMIR and ICON, which adopt the estimated SMPL, face the common problems of HMR-SMPL-based reconstruction methods–the predicted SMPL body bends legs or hunches over due to the depth ambiguity–leading to large reconstruction error. Similarly, PIFu and PIFuHD tend to have forward heads with slightly bending legs as they are not aware of any human body priors. One potential research direction is to *jointly train or optimize body prior (*e.g.*, SMPL, keypoints, human parsing) with mesh reconstruction*.

3) Ambiguity of geometry and appearance. As shown in the 6th-8th rows of Figure 3, the clothing prints can affect the surface reconstruction due to the ambiguity of geometry and appearance in single-view rendering. Among all these baselines, ICON and PIFuHD show high robustness to the input clothing prints as they predict normal images as intermediate representations, which reduces this ambiguity compared to directly inputting RGB images. Note that ICON only takes the normal images as input to the reconstruction module without using the RGB images, further mitigating the ambiguity and bringing even higher robustness than PIFuHD (the 8th example). Motivated by this observation, future research could shed more light on *disentangling the geometry and appearance either implicitly [37] or explicitly [7]*. Also, *predicting more comprehensive intermediate 2D/3D representations (*e.g.*, depth, illumination, keypoints, segmentation)* may also improve the performance.

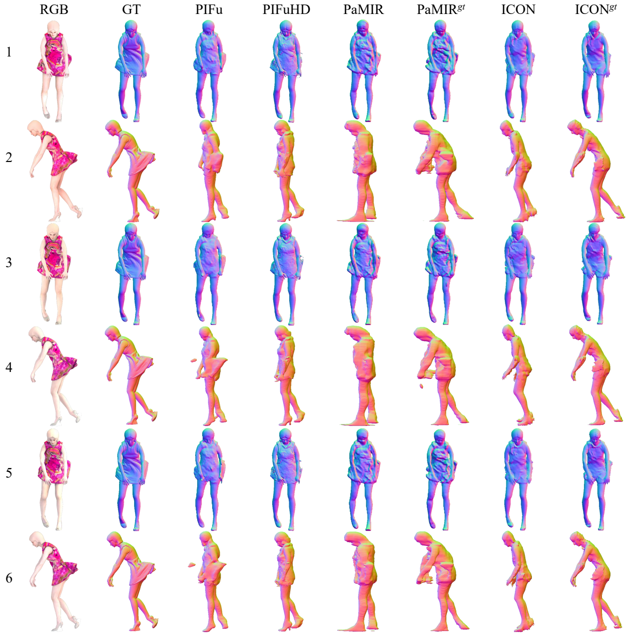


Figure 4. Temporal qualitative results. The 1st, 3rd, and 5th rows are three consecutive frames, and the 2nd, 4th, and 6th rows are the side views predicted from the corresponding front-view RGB. Refer to the supplementary material (Video B) for video results.

4) Temporal consistency. For real-time applications, *e.g.*, streaming from a monocular camera and importing the reconstructed motion sequence into a virtual scene, the temporal coherence of the generated meshes over time is vital for a high-quality user experience. We show the reconstructed meshes for three consecutive frames in Figure 4. As can be found from the side views of the reconstructed meshes, although only subtle motions are present in these frames, the HMR-SMPL-guided methods (PaMIR and ICON) suffer from unstable SMPL predictions and generate temporally inconsistent meshes. On the other hand, pure pixel-aligned methods (PIFu and PIFuHD) fail to predict accurate human pose but produce temporally consistent errors, thus having small ddt values. Given the ground truth SMPL meshes, PaMIR^{gt} is still more sensitive to global pose than ICON^{gt} as also noted by [49]. These observations are consistent with the quantitative temporal metrics reported in Table 2. It would be interesting to investigate *temporal modeling of implicit functions (*e.g.*, incorporating a recurrent neural network [27, 32, 55] to train the implicit function on the 4D dataset, or applying test-time fitting to refine the reconstructed meshes with temporal loss terms)*. We make such an attempt by adding a temporal term (penalize Chamfer distance between two successive frames) to the refinement process of PIFu$_{clo}$ and achieve better temporal consistency (Chamfer$_{ddt}$: 0.157→0.123 and Chamfer$_{dtd}$:

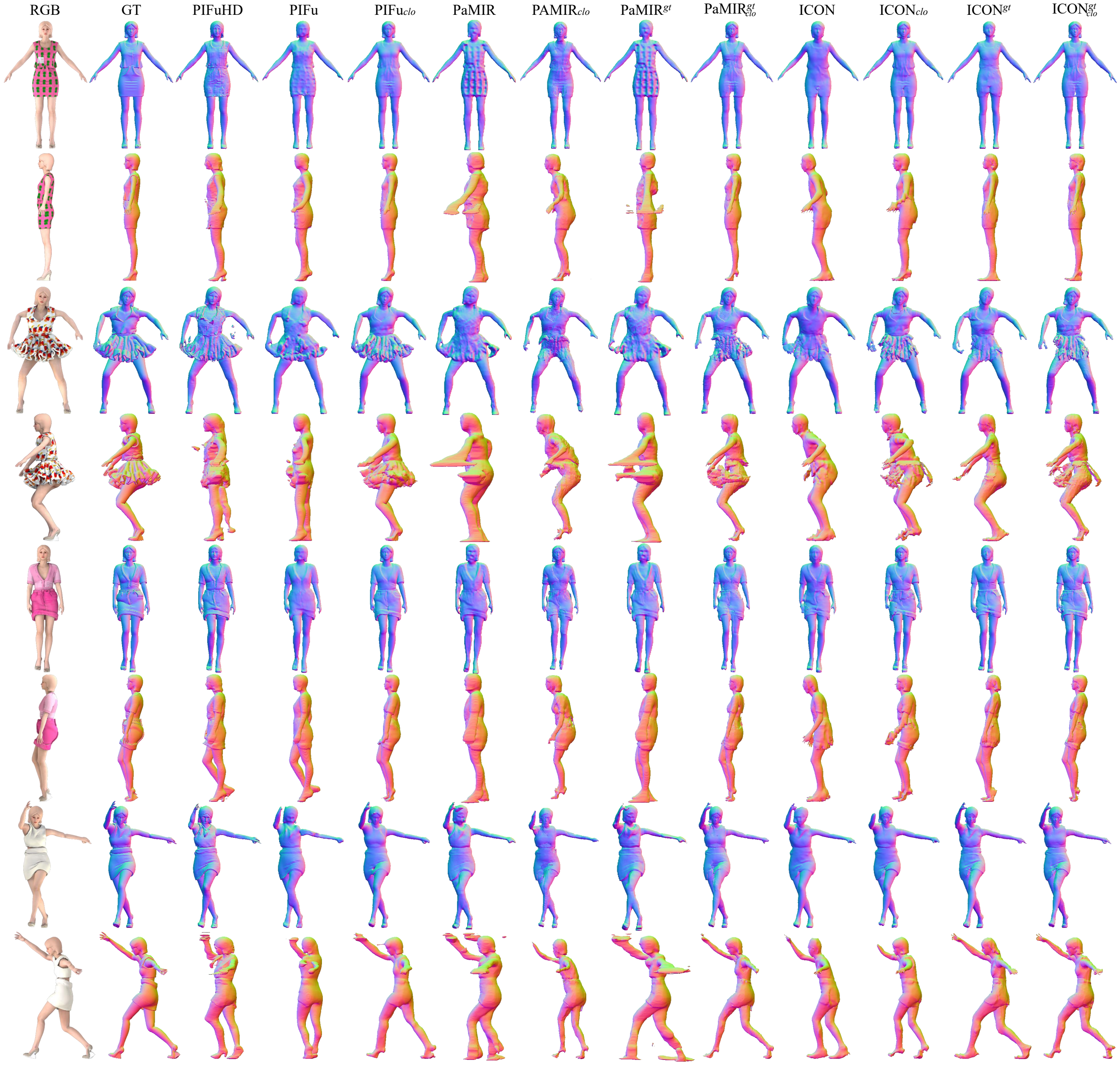


Figure 5. Qualitative results of baseline methods tested on CLOTH4D with front view and side view.

0.546→0.454).

4.4. Baseline Enhancement

In addition to the findings and insights discovered in Section 4.3 by evaluating on CLOTH4D. CLOTH4D, containing various clothing types and motion sequences, as well as true-to-life multi-layer thin 3D clothing structures, poses new challenges to clothed human reconstruction research. To investigate the SOTA performance when trained on CLOTH4D, we re-train PIFu, PaMIR, and ICON as described in Section 4.1, denoting as PIFu$_{clo}$, PaMIR$_{clo}$, and ICON$_{clo}$, respectively. We report the results of models trained on CLOTH4D in the gray columns in Table 2 and Figure 5. The following findings can be made:

1) In terms of quantitative metrics, the models trained on CLOTH4D generally outperform the original models trained on scan datasets as the data distributions of the training and testing datasets are closer. ICON$^{gt}_{clo}$ achieves the highest accuracy on the CLOTH4D test set. Furthermore, the qualitative results show that more high-frequency details are generated after training on CLOTH4D.

2) As the re-implementation setting in ICON's code allows the normal image to be input into all methods, the influence of garment print to PIFu and PaMIR are slightly relieved, validating the hypothesis we made in Section 4.3 that predicting intermediate representations like normals can re-

duce the ambiguity of geometry and appearance.

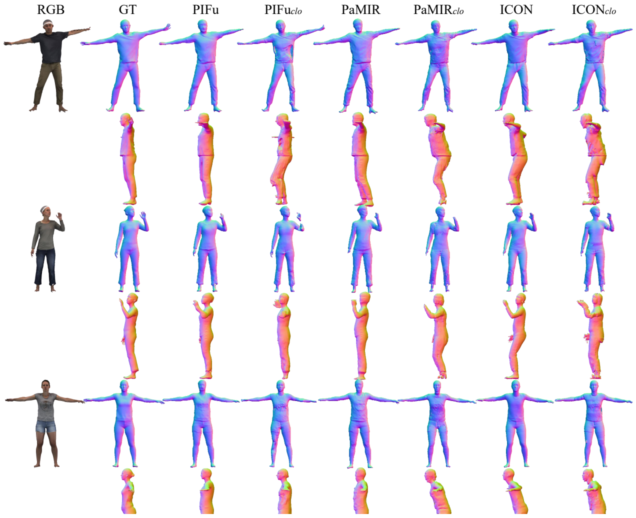
3) SOTAs fail to model layered and thin clothing structures as shown in the dress and skirt regions in Figure 5, where holes and tattered pieces are generated. Notably, the original PIFu and PaMIR can roughly generate the overall shapes of loose clothing, but they fail when trained on CLOTH4D. Since SOTA methods learn an occupancy field by sampling query points in the 3D space, for PIFu, whose spatial feature resolution is low, and for PaMIR, whose volumetric feature space is also of low resolution, it is hard to sample informative points near the thin surface. It is even harder for the network to learn if a querying point is inside or outside the mesh near the thin structure, as the inside and outside samples have very similar local features.

4) The difficulties of learning the occupancy field for thin faces motivates *developing methods that focus more on the query points near the thin surface* for better reconstruction. Future research may also *seek better implicit representations to boost the performance of reconstructing multi-layer thin structures e.g.*, [12, 13]. However, as shown in the supplementary material (Figure B), the state-of-the-art implicit representation cannot achieve satisfactory multi-layer thin structure reconstruction even if we feed the ground truth mesh as the input to the implicit function.

As shown in Table 3 and Figure 6, the performance on the CAPE dataset drops after training on CLOTH4D due to two reasons. 1) The CAPE dataset is collected in a controlled lab environment with dim lighting, which further enlarges the domain gap between CAPE and CLOTH4D. Consequently, the predicted normal images and reconstructed meshes are less accurate. 2) CAPE dataset contains single-layer scan meshes and tight clothing. However, models trained on the multi-layer CLOTH4D dataset tend to generate gaps between the clothing and skin, resulting in holes in the clothing and unsmooth surfaces. This also verifies that local features are sensitive to overfitting. It remains a challenging but interesting problem to have *a unified representation of single-layer and multi-layer data*, thus different types of datasets could be trained together to yield better generalizability. Finally, we show the reconstructed samples of the in-the-wild images in Figure 7.

4.5. Limitations

Firstly, the simulation results are generated via graphics software, which may crash in some cases. We show some detailed examples in the supplementary material (Figure C). Meanwhile, the results on CAPE show that the clothing features of current CLOTH4D can well represent basic men's wear. However, considering the completeness, the dataset scale and diversity of subjects will be further improved to cover males and kids. In addition, the original simulated mesh sequences are non-watertight, which must be converted to watertight (and such conversion is usually lossy) before feeding into current clothed human reconstruction methods. It would be essential to conduct future research that reconstructs meshes with arbitrary topologies. Finally, as indicated by the results on CAPE, the generalization ability to scan data is yet to be further explored.



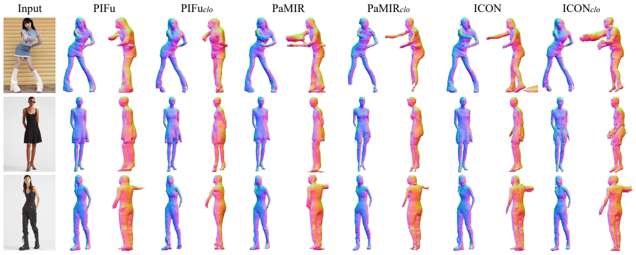


Figure 7. In-the-wild front and side view reconstructions. SOTAs with implicit functions tend to generate broken results.

5. Conclusion

We introduce CLOTH4D containing realistic and rich categories of clothing, avatars, and animations, and will release it for free, hoping to push the research on clothed human reconstruction. We evaluate current SOTAs with the newly introduced temporally-aware metrics and in-depth analyze their pros and cons by leveraging the advantages of CLOTH4D. We retrain those SOTAs on CLOTH4D, discuss the challenges the new dataset brings, and propose potential research directions. Although layering clothing in CLOTH4D brings immense difficulties to current research, we believe it is an important step toward more realistic and temporally coherent clothed human reconstruction.

Acknowledgement: This work is supported by Laboratory for Artificial Intelligence in Design (Project Code: RP 3-1) under InnoHK Research Clusters, Hong Kong SAR Government.

References

[1] 3dpeople. https://www.3dpeople.com. 2

[2] Adobe system incorporated. https://www.mixamo.com. 4

[3] Clo virtual fashion llc. https://www.clo3d.com. 2

[4] Official code of icon. https://github.com/YuliangXiu/ICON. 4

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 3

[6] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 3

[7] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 3, 4, 6

[8] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 2, 3, 4

[9] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 3

[10] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. *arXiv preprint arXiv:2204.13686*, 2022. 3

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[12] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022. 8

[13] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 8

[14] Mikhail Erofeev, Yury Gitman, Dmitriy S Vatolin, Alexey Fedorov, and Jue Wang. Perceptually motivated benchmark for video matting. In *BMVC*, pages 99–1, 2015. 4, 5

[15] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. *arXiv preprint arXiv:2206.02194*, 2022. 3, 4, 5

[16] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2, 3

[17] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 2, 3

[18] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019. 4

[19] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 4

[20] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 3

[21] 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*, pages 11046–11056, 2021. 3, 4

[22] Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision – ECCV 2020*, pages 512–530. Springer International Publishing, 2020. 3

[23] Jingwei Huang, Yichao Zhou, and Leonidas Guibas. Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. *arXiv preprint arXiv:2005.11621*, 2020. 4

[24] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3

[26] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 3

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 6

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[29] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 4

[30] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 667–684, 2018. 2

[31] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 3

[32] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 4, 6

[33] 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(6):1–16, 2015. 2, 3

[34] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[35] 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 Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2, 4

[36] Meysam Madadi, Hugo Bertiche, Wafa Bouzouita, Isabelle Guyon, and Sergio Escalera. Learning cloth dynamics: 3d + texture garment reconstruction benchmark. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR*, volume 133, pages 57–76, 2021. 2, 3

[37] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 6

[38] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2, 3

[39] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 3

[40] Albert Pumarola, Jordi Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2019. 2

[41] 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 International Conference on Computer Vision*, pages 2304–2314, 2019. 3, 4

[42] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 3, 4

[43] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, pages 355–366. Wiley Online Library, 2019. 2, 3

[44] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 3

[45] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 527–536, 2002. 4

[46] Matthew Trumble, Andrew Gilbert, Charles Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, 2017. 3

[47] Tuanfeng Y Wang, Duygu Ceylan, Jovan Popovic, and Niloy J Mitra. Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335*, 2018. 2, 3

[48] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 3

[49] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 3, 4, 5, 6

[50] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016. 4

[51] 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)*, June 2021. 2

[52] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2

[53] 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. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 4

[54] 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, 44(6):3170–3184, 2021. 3, 4

[55] Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, and Edmond Boyer. Spatio-temporal human shape completion with implicit function networks. In *2021 International Conference on 3D Vision (3DV)*, pages 669–678. IEEE, 2021. 6

[56] Xingxing Zou, Xiangheng Kong, Waikeung Wong, Congde Wang, Yuguang Liu, and Yang Cao. Fashionai: A hierarchical dataset for fashion understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 4

[57] Xingxing Zou and Waikeung Wong. fashion after fashion: A report of ai in fashion. *arXiv preprint arXiv:2105.03050*, 2021. 4