4DHumanOutfit: a multi-subject 4D dataset of human motion sequences in varying outfits exhibiting large displacements

Matthieu Armando, Laurence Boissieux, Edmond Boyer, Jean-Sébastien Franco, Martin Humenberger, Christophe Legras, Vincent Leroy, Mathieu Marsot, Julien Pansiot, Sergi Pujades, et al.

4DHumanOutfit: a multi-subject 4D dataset of human motion sequences in varying outfits exhibiting large displacements

Matthieu Armando[1] Laurence Boissieux[2] Edmond Boyer[2]
Jean-Sébastien Franco[2] Martin Humenberger[1] Christophe Legras[1]
Vincent Leroy[1] Mathieu Marsot[2] Julien Pansiot[2] Sergi Pujades[2]
Rim Rekik[2] Grégory Rogez[1] Anilkumar Swamy[1,2] Stefanie Wuhrer[2]

Figure 1: Identity and outfit axes of the 4DHumanOutfit dataset. 20 actors were captured in 7 outfits each while performing 11 motions per outfit. The figure shows the identities and the outfits of the subset that we release, with male actors on the left and female actors on the right. First row shows minimal clothing, which we leverage to obtain body shape parameters by fitting a parametric body model.

Abstract

This work presents 4DHumanOutfit, a new dataset of densely sampled spatio-temporal 4D human motion data of different actors, outfits and motions. The dataset is designed to contain different actors wearing different outfits while performing different motions in each outfit. In this way, the dataset can be seen as a cube of data containing 4D motion sequences along 3 axes with identity, outfit and motion. This rich dataset has numerous potential applications for the processing and creation of digital humans, e.g. augmented reality, avatar creation and virtual try on. 4DHumanOutfit is released for research purposes at `https://kinovis.inria.fr/4dhumanoutfit/`. In addition to image data and 4D reconstructions, the dataset includes reference solutions for each axis. We present independent baselines along each axis that demonstrate the value of these reference solutions for evaluation tasks.

1 Introduction

4DHumanOutfit is a new dataset of 4D human motion sequences, sampled densely in space and time, with 20 actors, dressed in 7 outfits each, and per-

[1]NAVER LABS Europe
[2]Inria centre at the University Grenoble Alpes
[3]Authors ordered alphabetically.

forming 11 motions exhibiting large displacements in each outfit. We designed 4DHumanOutfit to enable the combined analysis of shape, outfit and motions with humans. This results in a dataset shaped as a cube of data containing 4D motion sequences with three different factors that vary along the axes identity, outfit, and motion. Fig. 1 illustrates the morphology and clothing axes of this cube.

Analyzing and modeling the dynamics of human garments during motion and across actors is a well-studied problem in computer vision and computer graphics, with the goals of understanding human motion from partial data and generating realistic digital human animations. This has applications in video understanding, including action and fashion recognition; telepresence, including virtual change rooms and fashion transfers; and entertainment, including animation content generation. Many existing works study this problem from a data-driven perspective, where the goal is to learn motion dynamics from example data. To facilitate these studies, three main types of datasets of humans in motion have been introduced. The first type of dataset contains 2D videos of dressed humans in motion e.g. [61], which allows to capture the appearance of rich clothing dynamics observed in real garments. More recently, large-scale 4D datasets of minimally clothed 3D human bodies in motion have been published, either captured using acquisition platforms e.g. [45] or computed by fitting models to sparse motion capture data e.g. [38], which allow to learn detailed 3D body shape deformations over time. To enhance such datasets with garments, recent works use physical simulation to drape clothing on this 4D data e.g. [51], enabling therefore to model clothing dynamics for synthetically generated garments. 4DHumanOutfit contributes 4D data sampled densely in space and time of human bodies captured in different outfits and different motions. This combines the advantages of existing 2D datasets of capturing dynamic behaviour of layered clothing, including complex dynamics caused by seams and friction, with the advantages of existing 4D datasets of containing 3D shape information, including fine-scale geometric details.

The data we present has been captured in a multi-view acquisition platform that acquires 68 synchronized RGB streams at 50 frames per second, which are subsequently used to reconstruct densely sampled 3D geometric models with texture information per frame. For this dataset, we provide the RGB videos with masked background, and the reconstructed motion sequences in 4D in different spatial resolutions.

To demonstrate the potential of our dataset, we perform a baseline evaluation along each of the three axis independently. To this end, we introduce three tasks together with evaluation protocols. For the identity axis, we aim to predict the body shape of an identity given a 4D motion sequence of an actor wearing an arbitrary outfit. As reference solution for evaluating this task, we provide sequences captured in minimal clothing with body shapes resulting from fitting a standard parametric human body model [35] to the data. For the outfit axis, we aim to retrieve the outfit in a standardized pose from a given 4D motion sequence of an actor wearing an arbitrary outfit. As a reference solution to evaluate this task, we provide static scans of the outfit acquired on a mannequin. For the motion axis, we aim to retarget motion between actors from a 4D motion sequence showing the source motion to a static 3D scan of the target actor. When applied to data within the datacube, reference solutions are available as every actor was acquired performing each of the motions in each of the outfits. These tasks demonstrate that each axis of the datacube provides unique information that can be exploited in a large variety of practical applications.

The main contributions of this work are:

- The introduction of 4DHumanOutfit, a datacube of dynamic 4D human motion of 20 actors in 7 outfits each, performing 11 motions in each outfit, i.e. 1540 sequences in total. A large subset of 18 actors in 6 outfits and 10 motions is released for research purposes.

- The proposition of associated evaluation protocols and reference solutions for three tasks, along the identity, outfit, and motion axes of the datacube.

2 Related Work

Capturing humans in clothing has attracted many efforts in computer vision and graphics. Existing works can be mainly clustered into datasets containing (i) 2D images of clothed persons, (ii) synthetic 3D models, and (iii) 3D scans of humans in motion, different identities and various outfits. We review them briefly to highlight how 4DHumanOutfit goes beyond the state of the art.

2.1 2D fashion datasets

Many works focus on the creation of datasets of 2D images, as these are relatively easy to acquire. This includes early works, such as the Fashionista dataset [63] or works targeting to describe clothing with semantic attributes [15], as well as more recent datasets, such as FashionPedia [27], among many others made available for research purposes. A recent survey [16] provides an exhaustive list of published datasets until 2020, with a comprehensive classification of the tasks that can be addressed with 2D data. These include landmark detection, clothing parsing, retrieval, and attribute recognition. In addition, these datasets have allowed to tackle the task of virtual try on, where one can create high quality, compelling images [34, 22] or videos [18] of how a person would look like wearing a given clothing in a target pose. While the generated images reach impressive realistic quality, their applicability is yet limited, as they cannot be used for actual metric fit assessment.

2.2 Synthetic datasets

Capturing data with cameras or 3D scanners usually requires tremendous human and material efforts. In order to circumvent this issue, generated synthetic datasets of clothed humans can be considered. In this category works have proposed different datasets, with one or multiple characters in different settings, by leveraging 3D editing tools, such as MakeHuman [3], Mixamo [4], or human body models, such as SMPL [35]. For example, several datasets such as SURREAL [57], MHOF [47] or LTSH[24], place 3D models of humans on background images. These datasets have been designed to address the task of 2D or 3D pose estimation from a single image. Other datsets, such as AGORA [43], have increased the challenge by including images of multiple dressed persons with plausible interactions with the environment. As all these images are static and lack 3D realism, they do not capture and model the complexity of real cloths' dynamics.

To include dynamics in the data, most works leverage physics simulators, which can account for the type of clothing through physical parameters. Since the early work of Guan et al. [21], several methods have considered different clothing parameters, which allow creating plausible variations in the wrinkle patterns present on cloths. Synthetic datasets, with modest sizes such as BCNet [28], or larger scale datasets, such as 3DPeople [46] and Cloth3D [8, 37] provide 3D models of clothed humans. The last two explicitly explore the three dimensions of identity, motion, and clothing. While the explored range of subjects, poses, and cloth variations is impressive, the realism of these data are limited by the accuracy of the simulator used to create the data. The proposed 4DHumanOutfit dataset takes an alternative approach by capturing reality in a multi-view studio.

Interesting recent works have even modeled synthetic clothing at the sewing pattern level, allowing to automatically adjust the garment size to a personalized shape [31]. In our 4DHumanOutfit dataset we also release scans of the clothes on mannequins, which could allow to work on the clothes at the *sewing pattern level*.

All these works providing synthetic datasets have highly contributed to the community to advance the algorithmic approaches. With our work we argue that the acquisition of actual humans performing dynamic motion in varied clothing is necessary to validate the applicability of existing approaches to real data.

2.3 Scanned 3D humans datasets

With 4DHumanOutfit we explore the three axis of motion, identity, and clothing. We briefly review existing datasets that consider similar axes.

Motion. Human pose plays a crucial role in many application fields, such as medicine, sports or graphics, thus it has attracted many research efforts.

Marker based motion capture. A classical approach to capture human motion is to use motion capture (MoCap) systems with e.g. reflective markers. Following the pioneering dataset HumanEva [52], many other datasets have been acquired: Human3.6M [26], Total Capture [56], AMASS [38] or HUMAN4D [14]. The reflective markers allow to extract a good approximation of the pose, which is considered ground truth. Other modalities, such as video, depth sensors or inertial sensors, are simultaneously acquired. From this paired data, researchers have studied how to infer a pose from these other modalities. In addition, massive datasets, such as AMASS [38], have allowed to learn human pose priors which are widely used in the literature. While yielding precise information on poses with the marker locations, MoCap systems provide only sparse information on motion and imply complex setups with markers to be placed on subjects.

Markerless approaches. Another strategy to capture the human pose and motion is to use markerless systems, relying for that purpose on monocular settings [41, 5, 62, 42, 61]; on passive multi-view video systems, like for instance HUMBI [66] and Hi4D [65], the PanopticStudio [53, 30]; or on active systems such as 3DMD, used to acquire for example DynamicFaust [11], Flame [33] or Mano [50]. Depth camera setups have also been used to capture 4D motion sequences of multiple subjects [12]. For our work we use the Kinovis acquisition platform [2], a passive multi-camera system with a wide acquisition volume that enables dynamic displacements of the subjects and rich clothing dynamics.

Identity. In the identity axis, the seminal CAESAR dataset [49], created to study body morphology and clothing sizing purposes, contains scans of over 4500 individuals in 3 static poses each. Further datasets have scanned different persons in static [6, 23, 10] or dynamic [45] situations. Hasler et al. [23] provide a total of 500 static poses of 114 individuals, while the FAUST dataset [10] contains 10 individuals in 10 static poses each. DYNA [45] contains 10 individuals in a total of 129 sequences, which have been accurately registered for benchmarking in the DynamicFAUST dataset [11]. All these datasets have allowed the study of identity and static or dynamic pose, but have not considered the clothing axis.

Outfit. Early efforts have focused on capturing and analyzing sequences of few individuals captured in a single outfit e.g. [54, 17, 59]. More recently, different tasks related to clothing have motivated the creation of additional 3D datasets of real clothed humans.

For example, to explore how different sizes of cloths drape on the same human, the dataset SIZER [55] captured around 2000 static scans, from 100 subjects wearing clothes of different sizes. As all subjects strike the same A-pose, this dataset does not allow to study cloth dynamics.

To tackle the problem of estimating the shape under clothing, Yang et al. [64] and Zhang et al. [67] acquired scans of different subjects, with and without clothing, performing several motions. These datasets consider 6 subjects, 3 motions, and 3 outfits for Yang et al. [64] and 5 subjects, 3, motions and 2 outfits where 4DHumanOutfit considers 20 subjects in 7 outfits and 11 motions. Other datasets have been acquired with consumer RGBD sensors [68], providing lower quality than 4DHumanOutfit.

To learn a generative model of clothing, the dataset CAPE [36] was released. It also includes scans and SMPL mesh registrations from the ClothCap work [44] and contains 15 subjects in 4 different outfits performing different motions. The 4DHumanOutfit dataset is larger with 20 subjects and 7 clothing styles with richer dynamics. In addition, the systematic acquisition of all subjects performing very similar motions in all outfits provides an unprecedented opportunity to study how dynamic cloth deformations behave depending on identity, motion, and clothing.

3 4DHumanOutfit Dataset

In the following we detail the acquisition setup, the constitution of a cohort of subjects, the selected out-

fits, motions, and their captures.

3.1 Data acquisition

Motion sequences were captured by 68 calibrated RGB cameras (4 megapixels, 50 frames per second, focal lengths between 8 and 16mm) positioned roughly on a half-ellipsoid with radii 4m and 5m and height 5m looking towards the stage centre, for an average image resolution of 2.5mm per pixel at the scene centre. The total capture area covers a length of 5.5m and a width of 3.5m. Fig. 2 shows the multi-camera platform [2].

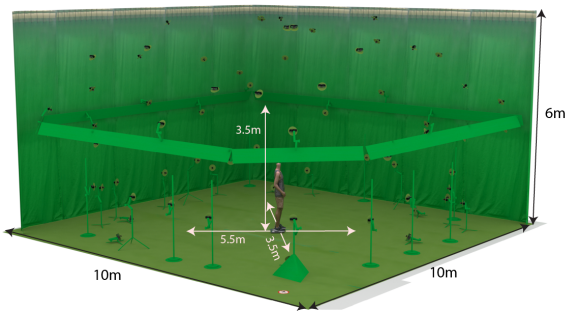

Figure 2: Multi-cameras acquisition platform [2] used to capture the 4DHumanOutfit dataset.

The capture and reconstruction pipelines, shown in Fig. 3 and 4, consist of several steps. First, synchronized video streams are acquired and silhouettes are segmented with the software [1] of the multi-camera platform [2]. Second, the resulting images and silhouettes are undistorted using the calibration information and masked with projections of inflated visual hulls. This significantly reduces the size of the data.

3D reconstructions are computed independently per frame, which results in a densely sampled 3D mesh per time instant. These meshes are obtained by performing multi-view reconstruction [32] on the undistorted and masked images. We decimate the resulting reconstruction into lower resolutions of 250k, 65k, 30k, and 15k vertices. The 3 lower resolution meshes are texture mapped using the capture platform software [1], which is not designed to handle higher resolutions.

3.2 Identities

10 females and 10 males were recorded. The participants were empirically chosen to cover the main variations of body shape according to eigenvectors computed on the CAESAR dataset [49]. The body shapes of all actors are shown in Fig. 5.

We release data of 9 female and 9 male actors, and keep the remaining data hidden to allow for future evaluations on unseen data.

3.3 Outfits

Motion recordings Each actor was recorded wearing their own arbitrary clothes and 6 additional outfits, chosen to cover a wide range of typical casual European clothing. The outfits differ in terms of their fit, from tight to wide, and are made of various materials, which results in rich dynamic behaviour during motion. Outfits are different for males and females. The following outfits, shown in Fig. 6, were used for women:

- **own** the actor's own clothes, unique to each actor, with the purpose to increase variability;
- **tig** socks, dotted white leggings, dotted salmon tank top, pink swimming cap (minimal clothing);
- **sho** white and pink sneakers, yellow shorts, purple T-shirt;
- **jea** yellow ballerinas, jeans, green and pink flowery shirt;
- **cos** yellow ballerinas, jeans, flowery purple dress;
- **opt** pink flip-flops or cream high heels, short grey dress or long loose blue dress or long tight red dress; as apparel for females offers more diversity than for males, we chose 3 optional outfits, using different materials and shapes to increase variability.
- **hidden** we recorded one additional outfit which will not be released to allow for future evaluations on unobserved data.

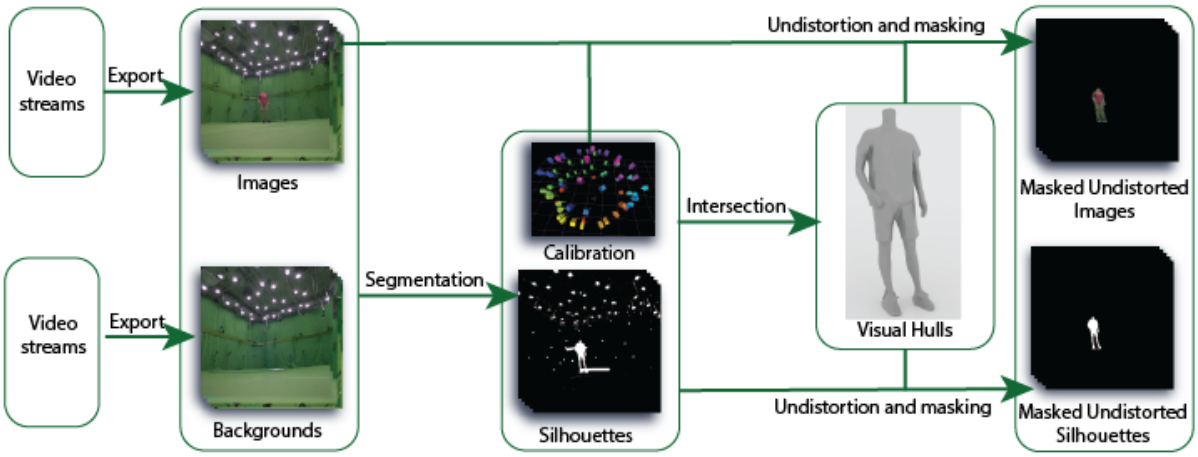

Figure 3: Data capture pipeline. Acquisition, production of visual hulls [1], and pre-processing of input images.

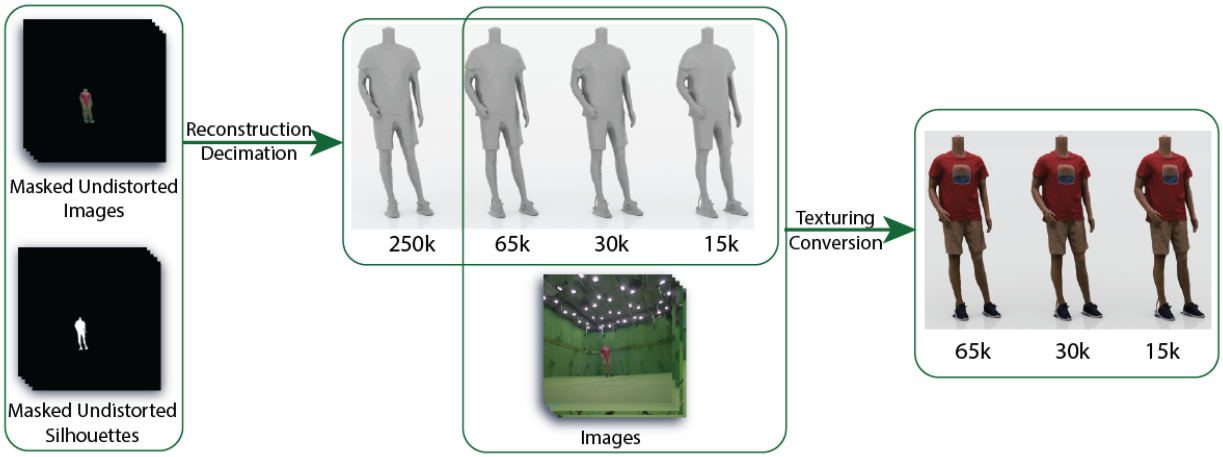

Figure 4: 3D reconstruction pipeline. Multi-view reconstruction [32] and final textured meshes.

For men, the following outfits, shown in Fig. 7, were used:

- **own** the actor's own clothes;
- **tig** socks, beige shorts, grey tank top, blue swimming cap (minimal clothing);
- **sho** blue and white sneakers, beige shorts, orange T-shirt with picture;
- **jea** black moccasins, jeans, grey and white striped shirt;
- **cos** black moccasins, dark costume trousers, grey and white striped shirt, dark costume jacket;

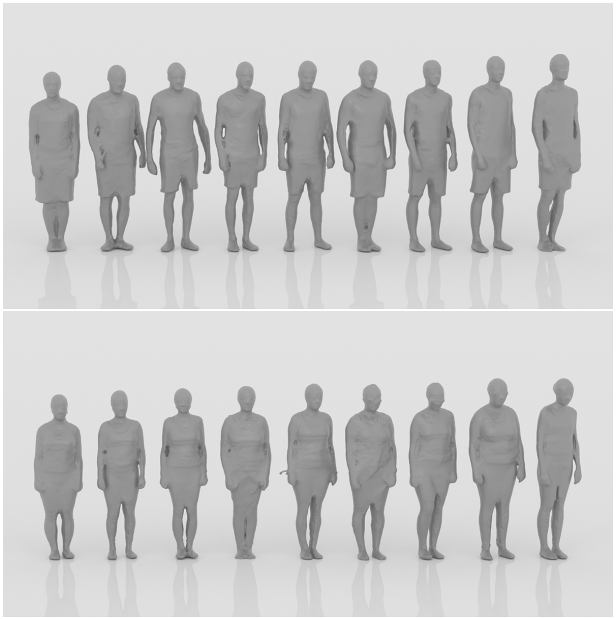

Figure 5: Morphologies. Men (top), women (bottom).

Figure 6: Female outfits. From left to right: own, tig, sho, jea, cos, opt1, opt2, opt3.

- **opt** black moccasins, dark costume trousers, grey and white striped shirt, beige trench coat;
- **hidden** we recorded one additional outfit which will not be released to allow for future evaluations on unobserved data.

Each actor was recorded in 7 outfits, including all non-optional ones and one optional outfit.

Figure 7: Male outfits. From left to right: own, tig, sho, jea, cos, opt.

Reference scans In addition to clothed human motion data, we acquired scans of each outfit. They can serve as reference solution for an outfit retrieval task. These models were acquired using two different systems as static scans of each outfit worn by a standard mannequin. The first scanning system used is our multi-camera platform; it was used to scan the mannequins without clothing and to record 8 scans for each outfit to allow for some natural variability in the clothing folds. The second scanning system is an Artex Eva structured light scanner, with scan resolution of about 1500k vertices.

Fig. 8 shows the female and male standard mannequins 65k reconstructions without clothing. Fig. 9 shows the same mannequins with all outfits. Fig. 10 shows the reconstructed male mannequin mesh at resolutions 250k and 1500k.

Figure 8: Scans of standard female and male mannequins meshes.

Figure 9: Outfits scanned on female (top) and male (bottom) mannequins.

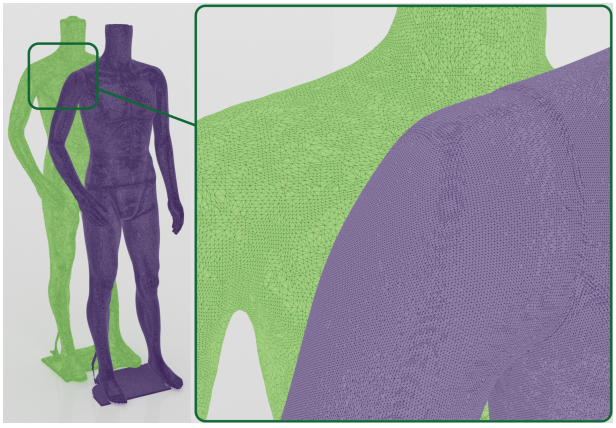

Figure 10: 250k reconstruction of the male mannequin (green hidden faces rendering) vs. 1500k scanned version (purple hidden faces rendering).

3.4 Motions

Actors were asked to perform 11 motions involving significant displacements within the scene. We focus on motions with large displacements as these are yet rare in existing captured 4D datasets. For instance DYNA [45] captures soft-tissue dynamics during motions but without large displacements. We further choose the 11 motions to contain variations in upper and lower body motions, while covering common motions including different variants of walking. The motions, illustrated in Fig. 11, are:

- **walk** a simple walk across the studio;
- **avoid** a walk with last-second obstacle avoidance;
- **back** a walk with a U-turn;
- **torso** a walk with a torso rotation to look backwards;
- **run** a jog / run across the studio;
- **jump** jump on the spot;
- **dance** a dance with both legs and arms wide motion;
- **hop** hopscotch;
- **2 free motions** to be chosen by the actor to increase the variability of the dataset, this included mostly martial art, dance and other sport motions;
- **hidden** we recorded one additional motion which will not be released to allow for future evaluations on unobserved data.

The duration of the recorded sequences ranges from 0.8 (for a free motion) to 17.2 (for dance motion) seconds.

3.5 Summary

A total of 1617 sequences were recorded, involving the processing of 459080 frames. The computations were handled on 2 clusters (17 16-core servers equipped with Nvidia Quadro 4000 cards and 20 16-core Intel Xeon CPU servers) resulting in the generation of 540TB of total data during the project. Fig. 12 provides an overview of the storage space required by

Figure 11: Representative frames of the motions. From left to right: walk, avoid, back, torso, run, jump, dance, hopscotch, free 1, free 2. Here free 1 is boxing, free 2 is kick. Men (top), women (bottom).

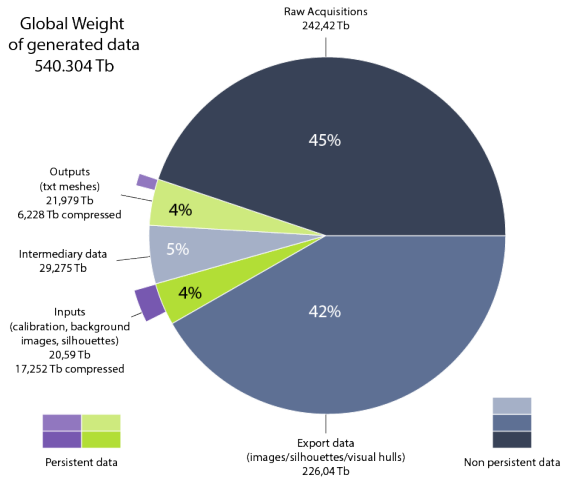

Figure 12: Storage volume required during data processing.

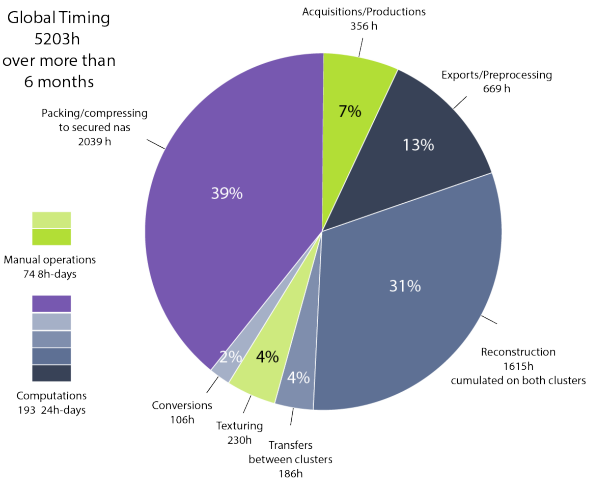

Figure 13: Computation time required during data processing.

the data during the generation of 4DHumanOutfit. The final 4D dataset consists of meshes in different resolutions (250k, 65k, 30k, 15k vertices), and undistorted and masked images and silhouettes. The total volume is 22TB, 20.5TB of which are occupied by the undistorted and masked images and silhouettes.

Concerning the timing, the project stretched over a bit more than 6 months, for about 5200 hours, divided as shown in Fig. 13. A significant amount of time was dedicated to packing and compressing.

4 Evaluation

The 4DHumanOutfit datacube allows to learn correlations between identity, outfit and motion. To demonstrate its potential, we perform a simple evaluation along each of the three factors independently. This demonstrates that each axis of the datacube provides unique information that can be exploited in practical applications.

4.1 Identity

The first evaluation aims to estimate the body shape of the identity performing a motion from an arbitrary sequence of the datacube. That is, given as input a 4D motion sequence showing a person (of arbitrary identity and outfit) in motion, the aim is to estimate the undressed body shape in T-pose. This problem has been studied previously in e.g. [64, 67], and is of interest for virtual change room applications.

Evaluation protocol To evaluate the accuracy of the retrieved identity in T-pose, we compute for each identity a reference solution of the naked body shape. This is achieved by fitting a parametric human body model to each frame of the person captured in minimal clothing, and by combining the resulting information. That is, we use the minimal clothing regime, which is very close to the actor's skin, as proxy for true body shape.

We use a parametric human body model with two sets of parameters, one representing body shape and one representing static pose. Changing the static pose parameters to the ones representing a T-pose allows to re-pose the body shape into standard pose. In practice, we use the SMPL model [35].

To reliably fit the body model to the sequences captured in minimal clothing, we use an existing framework [25] based on SMPLify [9]. For a given timestep, we compute 2D keypoints in all images with alphapose [19], and optimize a SMPL body model with respect to said keypoints, with an additional loss to force the reprojection of the SMPL mesh to fit inside the silhouette on all images, and the pose and shape priors used in SMPLify.

This is done for all sequences captured in minimal clothing. We then select a fit that minimizes the Chamfer distance to the corresponding 3D reconstruction. The resulting body shape parameters are used to reconstruct a model in T-pose, which is used as reference body shape. For each identity, its reference body shape is released along with the dataset.

To quantitatively evaluate the quality of a body shape estimate computed from a dressed 4D motion sequence of identity i, we compute the Chamfer distance between the result and the reference body shape of i, in a standard T-pose.

Baseline As a baseline method for estimating body-shape parameters, we use the same optimization method described above, based on keypoints and silhouettes, but applied directly to dressed sequences. The setting is more complex, as estimated keypoints are generally less precise on frames with loose clothing, and silhouettes are farther from the silhouette of the body shape. This baseline is computed per-frame.

Results We give here numerical and qualitative results of the baseline. The optimization is sensitive to the input keypoints, so results tend to be noisy. When computing the reference shape, this problem is addressed by using the 3D reconstruction to select the best fit. However, the results of the baseline shape estimation are affected by this problem. It sometimes leads to overly small or elongated shapes, when the optimization does not converge to a good solution. In cases of convergence, estimated body shapes are often too big, as the loose clothing is larger than the body shape.

Fig. 14 shows color-coded results for female **opt** and male **cos** outfits. In these examples, the body shape estimates computed by our baseline are close to the reference shape in terms of height and overall body shape. However, the volume of the body shape is overestimated in areas that are occluded by clothing. The reason is that the baseline fits the largest body model that can fit in the silhouettes, and does so on a per-frame basis.

Limitations For privacy concerns, we use sequences in tight clothing to compute reference shapes. While this causes small errors due to clothing folds and the width of the cloth, the resulting error is small compared to typical human motions.

We chose a simple baseline to illustrate the challenges of this task. It could be improved by taking information from the full sequence into account.

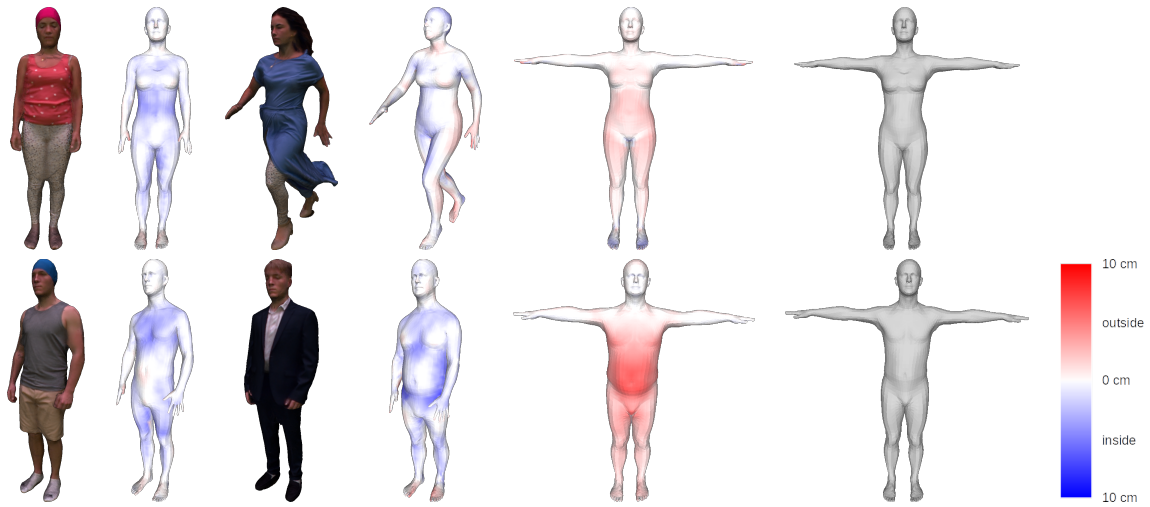

Figure 14: Identity evaluation. Given a sequence in minimal clothing (a), the naked body shape is estimated. This allows to estimate a *reference* body shape for evaluation (b). Vertices are color-coded according to the distance to the reconstructed surface (a). Given an arbitrary clothed sequence as input (c), the naked body shape is estimated (d). Reposing both this result (e) and the reference shape (f) into T-pose allows for quantitative comparison, shown as color-coding.

4.2 Outfit

The second evaluation aims to retrieve information related to outfit. Let $J^{3d}_{\mathcal{M}}$ denote the pose of the mannequin shown in Fig. 8. Given as input a 4D motion sequence showing a dressed person in motion, the aim is to retrieve the outfit in pose $J^{3d}_{\mathcal{M}}$. This way of evaluating the outfit independently of identity and dynamics is novel to our knowledge. The related problem of inverse cloth design, where the goal is to retrieve a rest pose unaffected by physical forces for use in a physical simulator given the shape of a garment, has been studied [13], but differs from our scenario as we are interested in retrieving the shape of the outfit in standard pose (as affected by physical forces).

Generating 3D garment deformations using physics-based simulators is challenging, because the generation of realistic detailed 3D garment models is typically done by trained artists and costly, and because physical simulators require input parameters that need to be tuned. 4DHumanOutfit contains accurate 3D garment deformations for a number of outfits, and has the potential to be used for data-driven garment synthesis without relying on physics-based simulators.

Evaluation protocol To evaluate the accuracy of the retrieved outfit in pose $J^{3d}_{\mathcal{M}}$, we use the reference scans acquired for each outfit draped on the mannequin as pseudo ground truth. The retrieved outfit is compared to its corresponding reference scan by measuring the Chamfer distance between the two shapes. In particular, we report the Chamfer distance in mm between the retrieved garment mesh $\mathcal{V}_{retr}$ and the reference scan $\mathcal{V}_{\mathcal{M}}$.

Baseline We propose a simple baseline to solve this problem by framing it as a retrieval task. We first compute the 3D joints of each frame in a 4D motion sequence by fitting SMPL to the models as described in the previous section, and then retrieve the frame whose pose J^{3d}_{t} is closest to $J^{3d}_{\mathcal{M}}$. The pose distance is computed by considering a Procrustes-aligned dis-

tance as

$$\arg\min_{J_t^{3d}} \left(\sum_i \mathrm{D}_P(J_t^{3d}, J_{\mathcal{M}}^{3d}) \right), \quad (1)$$

where D_P is the distance in joint angle space. The 3D joints on the mannequin are obtained by manually annotating 3D points on the surface of the mannequin scan.

Results Results are analyzed for three different subjects wearing three different outfits. Fig. 15 shows three viewpoints of the retrieved garment and the reference scan for each of the three subjects. Note that the simple baseline already retrieves poses with garments that have visually similar wrinkle patterns. Tab. 1 reports the Procrustes-aligned distance from Eq. 1. The error ranges from $26mm$ to $64mm$ across different subjects. This error has high variance because the retrieved outfit is one frame of the input sequence. If the input 4D motion does not contain frames in a pose similar to $J_{\mathcal{M}}^{3d}$, the error is high. Furthermore, the reference scan does not contain head surface, and the mannequin's body shape may not be close to the body shape of the input sequences.

Limitations We propose a novel outfit retrieval task that has potential applications in online garment retrieval. The task along with the reference solutions that allow for quantitative evaluation have the potential to allow for further research in garment retrieval.

A major limitation of our protocol is that dynamic effects and body shape changes are currently not considered. That is, we frame the problem as a static one even though dynamic motion is present in the 4D motion sequence, and we ignore the influence of the wearer's body shape on the outfit geometry. The baseline we propose is simple and leaves significant room for improvement. However, it already demonstrates that outfit configurations visually similar to the reference scans can be found.

4.3 Motion

The third evaluation aims to examine the motion axis. Our evaluation considers the task of motion

	$\mathcal{V}_{retr} \rightarrow \mathcal{V}_{\mathcal{M}}$	$\mathcal{V}_{\mathcal{M}} \rightarrow \mathcal{V}_{retr}$
tig	63.8	60.1
sho	34.0	30.6
cos	26.0	29.8

Table 1: Quantitative outfit evaluation. Chamfer distance (in mm) between retrieved garment mesh $\mathcal{V}_{retr}$ and $\mathcal{V}_{\mathcal{M}}$ for outfits **tig**, **sho** and **cos**.

Figure 15: Outfit evaluation. Qualitative visualization of the retrieved outfit. Retrieved outfits and reference scans shown for 3 different outfits (left to right: tig, sho and cos). Three different viewpoints are shown.

retargeting where the objective is to generate a 4D motion sequence of a given identity that performs the same motion as another given 4D motion sequence. In particular, given as input the 4D motion sequence showing identity i_1 in outfit o performing motion m along with a 3D model of identity i_2 wearing minimal clothing, the goal is to compute the 4D sequence showing identity i_2 while performing motion m.

A challenge when evaluating motion retargeting is the lack of realistic ground truth data. On the one hand, realistic 4D ground truth is lacking due to the sparse nature of existing large 4D human datasets e.g. [38, 12] where all actors are not seen performing all motions. This lack of data encourages state of the art to evaluate on synthetically generated 3D motions. These synthetic motions are often generated with skinning methods and lack realistic local dynamic details. On the other hand, smaller 4D datasets [11, 45] with dynamic details do not account for clothing.

In the following, we show that 4DHumanOutfit can be leveraged to evaluate motion retargeting methods by providing captured reference solutions with accurate geometric details.

Evaluation protocol To evaluate the accuracy of the retargeted motion, we compare the 4D motion resulting from the retargeting to the target motion of identity i_2 in minimal clothing performing motion m captured in 4DHumanOutfit.

In this scenario, the retargeted motion $M_1 = \{m_{1,j}\}_{j=1}^{n}$ and the target motion $M_2 = \{m_{2,j}\}_{j=1}^{m}$ are characterized by sequences of 3D point clouds. The point clouds are not in correspondence, so we use the Chamfer distance to compare them. To compare M_1 and M_2, the Chamfer distance relies on a nearest neighbor search per point, which is heavily influenced by small variations of the global trajectory and temporal unfolding of M_1 and M_2. We remove this influence by spatio-temporally aligning M_1 and M_2, as this is common when evaluating retargeting approaches [39, 29].

To align the global trajectories, we center the pointclouds using their centroid c. To align the temporal unfolding of the motions, we use Dynamic Time Warping (DTW) [7]. Given two sequences of point clouds, DTW computes the optimal monotonic path p^* between aligned frames as

$$p^* = \arg\min_{p} \left(\sum_{j} \mathrm{D}_{Ch}(m_{1,j} - c_{1,j}, m_{2,p[j]} - c_{2,p[j]}) \right),$$

where D_{Ch} is the Chamfer distance. The proposed metric is then evaluated as the median error along this path as

$$\underset{j}{med} \left(\mathrm{D}_{Ch} \left(m_{1,j} - c_{1,j}, m_{2,p^*[j]} - c_{2,p^*[j]} \right) \right). \quad (2)$$

Baseline Motion retargeting has been approached from different angles, either using deformation models that directly operate on the body surface [60, 48], using structured latent representations of 4D motion [29, 40] or using skeletal representations which are linked to the body surface using an animation model [58, 39].

Some of these works require correspondences between the target and source bodies or temporal correspondences in the source motion. Our source motion is unregistered and we can leverage the template fitting from Sec. 4.1 to have access to a target identity in minimal clothing in T-pose. From the applicable methods [39, 60, 29], we choose [39] as our baseline because it generalizes well under various motion and shape preservation metrics and was already tested on the raw data of a multi-view acquisition setup. The baseline operates in three steps. First, the source skeleton is extracted using a PointFormer network. Second, this skeletal motion is retargeted to the target at the skeletal level using a recurrent network. Third, the dense geometry of the target shape is recovered using a learnt skinning prior.

Results Tab. 2 reports the metric of Eq. 2 for 4 retargetings considering 2 identities: a female and a male identity for 2 source motions and 2 source outfits. As the per sequence median error is more informative when comparing different methods, we also visualize the spatio-temporal distribution of the error for two retargetings to differentiate error due to the natural variability and the error introduced by the retargeting method in Fig. 16.

Fig. 16 shows the retargeting from male to female on a jumping motion and from female to male on

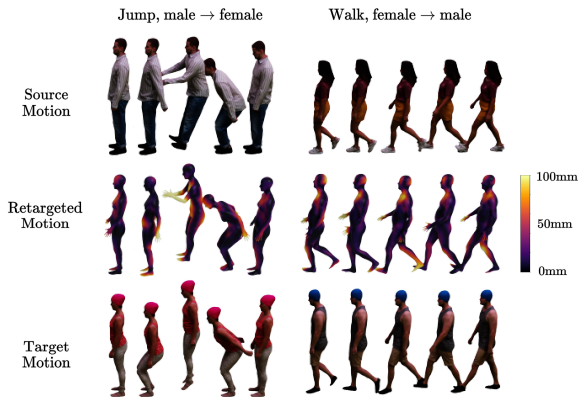

Figure 16: Motion evaluation. Qualitative visualization of the Chamfer distance for two retargetings. Top: Source motion with outfit. Middle: Retargeted motion, color coded by Chamfer distance to the target motion. Bottom: Target motion in tight clothing.

a walking motion. The color coding shows that the method generates plausible poses overall with large error (yellow color) due to natural variability in the arm pose between the source and target jumping motions. It also highlights that the method could be improved in terms of head and arm pose transfer (red color) with the head facing down in the jumping retargeting and incorrect arm poses in some frames of the walking retargeting.

	male,jea → female	female ,sho → male
walk	0.020	0.026
jump	0.026	0.035

Table 2: Motion evaluation. Median Chamfer distance (Eq. 2 in m) for 4 retargeting examples.

Limitations It is known that a fixed type of motion performed by the same performer is performed slightly differently at different trials [20]. This variability is not implemented in our baseline, which instead outputs a deterministic retargeting solution. To address this limitation, our evaluation protocol normalizes global trajectory and temporal unfolding.

Second, existing retargeting baselines that operate on raw scan data are limited to outfits that are close to the body surface. Hence, we cannot leverage more ample outfits present in 4DHumanOutfit.

5 Conclusion

We presented 4DHumanOutfit, a large-scale dataset of 20 actors wearing 7 outfits each and performing 11 motions per outfit. This data captures detailed spatio-temporal dynamics of varying outfits, and their interaction with different morphologies and motions. We demonstrated that each axis of the resulting data-cube contains unique information using simple evaluations. This data has the potential of serving in many different applications involving digital humans including augmented or virtual reality applications (e.g. virtual change rooms), and in entertainment (e.g. animation content generation).

6 Acknowledgments

This work was supported by French government funding managed by the National Research Agency under grants ANR-21-ESRE-0030 (CONTINUUM), ANR-19-CE23-0013 (3DMOVE), and ANR-19-CE23-0020 (Human4D).

References

[1] 4DViews 4DVManager software. Online `https://www.4dviews.com/`, 2023. Accessed on March 21st 2023.

[2] Kinovis, Inria 4D modeling multi cameras platform. Online `http://kinovis.inria.fr/inria-platform`, 2023. Accessed on March 21st 2023.

[3] Makehuman. Online `http://makehuman.org/`, 2023. Accessed on March 21st 2023.

[4] Mixamo. Online `https://www.mixamo.com`, 2023. Accessed on March 21st 2023.

[5] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005.

[7] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *International Conference on Knowledge Discovery and Data Mining*, page 359–370, 1994.

[8] H. Bertiche, M. Madadi, and S. Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359, 2020.

[9] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578, 2016.

[10] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

[11] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *Conference on Computer Vision and Pattern Recognition*, pages 5573–5582, 2017.

[12] Z. Cai, D. Ren, A. Zeng, Z. Lin, T. Yu, W. Wang, X. Fan, Y. Gao, Y. Yu, L. Pan, F. Hong, M. Zhang, C. C. Loy, L. Yang, and Z. Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577, 2022.

[13] R. Casati, G. Daviet, and F. Bertails-Descoubes. Inverse Elastic Cloth Design with Contact and Friction. Research report, Inria Grenoble, HAL id hal-01309617, 2016.

[14] A. Chatzitofis, L. Saroglou, P. Boutis, P. Drakoulis, N. Zioulis, S. Subramanyam, B. Kevelham, C. Charbonnier, P. Cesar, D. Zarpalas, et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020.

[15] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*, pages 609–623, 2012.

[16] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys*, 54(4):1–41, 2021.

[17] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3):#98,1–10, 2008.

[18] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *International Conference on Computer Vision*, pages 1161–1170, 2019.

[19] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2023.

[20] S. Ghorbani, C. Wloka, A. Etemad, M. A. Brubaker, and N. F. Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In *Computer Graphics Forum*, volume 39, pages 225–239, 2020.

[21] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *Transactions on Graphics*, 31(4):1–10, 2012.

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

[23] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 2(28):337–346, 2009.

[24] D. T. Hoffmann, D. Tzionas, M. J. Black, and S. Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition*, pages 609–623, 2019.

[25] B. Huang. Mvsmplfitting. `https://github.com/boycehbz/MvSMPLfitting`.

[26] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[27] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European Conference on Computer Vision*, pages 316–332, 2020.

[28] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35, 2020.

[29] B. Jiang, Y. Zhang, X. Wei, X. Xue, and Y. Fu. H4D: human 4d modeling by learning neural compositional representation. In *Conference on Computer Vision and Pattern Recognition*, pages 19355–19365, 2022.

[30] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2017.

[31] M. Korosteleva and S.-H. Lee. Generating datasets of 3d garments with sewing patterns. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[32] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *IEEE,International Conference on Computer Vision*, pages 3113–3122, 2017.

[33] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics*, 36(6):194:1–194:17, 2017.

[34] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

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

[36] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black. Learning to dress 3d people in generative clothing. In *Conference on Computer Vision and Pattern Recognition*, pages 6468–6477, 2020.

[37] M. Madadi, H. Bertiche, W. Bouzouita, I. Guyon, and S. Escalera. Learning cloth dynamics: 3d+ texture garment reconstruction benchmark. In *Conference on Neural Information Processing Systems Competition and Demos*, pages 57–76, 2020.

[38] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019.

[39] M. Marsot, R. Rekik, S. Wuhrer, J.-S. Franco, and A.-H. Olivier. Correspondence-free online human motion retargeting. *arXiv 2302.00556*, 2023.

[40] M. Marsot, S. Wuhrer, J.-S. Franco, and S. Durocher. A structured latent space for human body motion generation. In *Conference on 3D Vision*, pages 557–566, 2022.

[41] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *Conference on 3D Vision*, pages 506–516, 2017.

[42] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Conference on 3D Vision*, pages 120–130, 2018.

[43] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black. Agora: Avatars in geography optimized for regression analysis. In *Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021.

[44] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *Transactions on Graphics*, 36(4), 2017.

[45] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. *Transactions on Graphics*, 34(4):120:1–120:14, 2015.

[46] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3Dpeople: Modeling the geometry of dressed humans. In *International Conference on Computer Vision*, pages 2242–2251, 2019.

[47] A. Ranjan, D. T. Hoffmann, D. Tzionas, S. Tang, J. Romero, and M. J. Black. Learning multi-human optical flow. *International Journal of Computer Vision*, 128:873–890, 2020.

[48] J. Regateiro and E. Boyer. Temporal shape transfer network for 3d human motion. In *Conference on 3D Vision*, pages 424–432, 2022.

[49] K. Robinette, H. Daanen, and E. Paquet. The CAESAR project: A 3-D surface anthropometry survey. In *Conference on 3D Digital Imaging and Modeling*, pages 180–186, 1999.

[50] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics*, 36(6):245:1–245:17, 2017.

[51] I. Santesteban, M. A. Otaduy, and D. Casas. Snug: Self-supervised neural dynamic garments. In *Conference on Computer Vision and Pattern Recognition*, pages 8130–8140, 2022.

[52] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4, 2010.

[53] T. Simon, H. Joo, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *Conference on Computer Vision and Pattern Recognition*, pages 4645–4653, 2017.

[54] J. Starck and A. Hilton. Surface capture for performance-based animation. *Computer Graphics and Applications*, 27(3):21–31, 2007.

[55] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision*, pages 1–18, 2020.

[56] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *British Machine Vision Conference*, pages 1–13, 2017.

[57] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[58] R. Villegas, D. Ceylan, A. Hertzmann, J. Yang, and J. Saito. Contact-aware retargeting of skinned motion. In *International Conference on Computer Vision*, pages 9700–9709, 2021.

[59] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *Transactions on Graphics*, 28(5):174:1–12, 2009.

[60] J. Wang, C. Wen, Y. Fu, H. Lin, T. Zou, X. Xue, and Y. Zhang. Neural pose transfer by spatially adaptive instance normalization. In *Conference on Computer Vision and Pattern Recognition*, pages 5831–5839, 2020.

[61] T. Y. Wang, D. Ceylan, and K. K. Singh. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In *International Conference on 3D Vision*, pages 268–277, 2021.

[62] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *Transactions on Graphics*, 37(2):1–15, 2018.

[63] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012.

[64] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454, 2016.

[65] Y. Yin, C. Guo, M. Kaufmann, J. Zarate, J. Song, and O. Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition*, pages 17016–17027, 2023.

[66] J. S. Yoon, Z. Yu, J. Park, and H. S. Park. Humbi: A large multiview dataset of human body expressions and benchmark challenge. *Transactions on Pattern Analysis and Machine Intelligence*, 45(1):623–640, 2021.

[67] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Conference on Computer Vision and Pattern Recognition*, pages 5484–5493, 2017.

[68] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *International Conference on Computer Vision*, pages 7739–7749, 2019.