ClothPose: A Real-world Benchmark for Visual Analysis of Garment Pose via An Indirect Recording Solution

Wenqiang Xu[*1,2], Wenxin Du[*1], Han Xue[1], Yutong Li[1], Ruolin Ye[3], Yan-Feng Wang[1], Cewu Lu[§1,2]
[1]Shanghai Jiao Tong University [2]Shanghai Qi Zhi institute [3]Cornell University
[1]{vinjohn, mnkmYuki, xiaoxiaoxh, davidliyutong, wangyanfeng, lucewu}@sjtu.edu.cn
[3]ry273@cornell.edu

Abstract

Garments are important and pervasive in daily life. However, visual analysis on them for pose estimation is challenging because it requires recovering the complete configurations of garments, which is difficult, if not impossible, to annotate in the real world. In this work, we propose a recording system, GarmentTwin, which can track garment poses in dynamic settings such as manipulation. GarmentTwin first collects garment models and RGB-D manipulation videos from the real world and then replays the manipulation process using physics-based animation. This way, we can obtain deformed garments with poses coarsely aligned with real-world observations. Finally, we adopt an optimization-based approach to fit the pose with real-world observations. We verify the fitting results quantitatively and qualitatively. With GarmentTwin, we construct a large-scale dataset named ClothPose, which consists of 30K RGB-D frames from 2400 video clips on 600 garments of 10 categories. We benchmark two tasks on the proposed ClothPose: non-rigid reconstruction and pose estimation. The experiments show that previous baseline methods struggle with highly large non-rigid deformation of manipulated garments. Therefore, we hope that the recording system and the dataset can facilitate research on pose estimation tasks on non-rigid objects. Datasets, models, and codes will be made publicly available.

1. Introduction

We manipulate garments every day: we fold our T-shirts, flatten our suits, and perform other tasks that involve interacting with garments. A vision system to reconstruct the complete configuration of the garment can be beneficial for downstream tasks such as object understanding, VR/AR, and robotic manipulation. The problem of reconstructing the garment configuration is defined as *garment pose estimation* in GarmentNets [9]. However, problems arise when dealing with real-world data. Given that a garment has a near-infinite degree of freedom, it is extremely hard to record its pose in the real world. Real-world garments can easily undergo large non-rigid deformation, making them hard to annotate. As a result, previous works are limited to annotate key points [6, 46] or feature lines [4] when using real-world data. Such simplifications hinder the development of research on visual garment understanding. A question still haunts the community: *Is it really impossible to measure the garment pose in the real world?*

Taking the lessons from capturing poses of rigid or articulated objects in the real world [18, 33], passive sensors such as QR-like fiducials or reflective markers are commonly used. These markers are particularly effective for rigid object parts, as they share the same pose transformation for all the vertices. With such rigid constraints, the pose of markers in occluded regions can be inferred from visible regions in observation. However, for deformable objects such as garments, these rigid constraints do not apply, making it challenging to infer the pose from visible regions. Thus, the passive sensors cannot be applied to garments.

As for active sensors, their accuracy is often insufficient to meet the requirement of having an error range smaller than the thickness of a garment. A typical garment has a thickness of $0.05 \sim 1cm$. Therefore, an active measuring sensor should have an error of at least $1cm$ to prevent one side of the garment from penetrating the other. However, even magnetic sensors, one of the most accurate localization techniques available, only have an accuracy of $2.6cm$ within a $30cm \times 30cm \times 30cm$ region [7]. For a more in-depth discussion of active sensors, please refer to Sec. 2.

As the sensor technology seems to need a long time to catch up with the necessary accuracy, we turn to an indirect method to measure garment poses. Instead of directly measuring the garment pose from a static state, our method

* indicates equal contributions.

§ Cewu Lu is the corresponding author, the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute.

measures it based on its dynamic movements. We draw inspiration from continuum mechanics that the deformation is influenced only by external forces applied on the object from a given configuration. Therefore, knowing the initial pose of garments and the operation to deform it allows us to determine the pose during the operation. In this way, we can estimate the garment pose even when it is in large deformation, as long as it is tracked and starts from a simple configuration. Based on the observation, we propose a new recording setup for garments, named **GarmentTwin**.

The workflow of GarmentTwin consists of five major procedures: (1) **Model collection**: We prepare 3D models of the real-world garments by scanning. (2) **Manipulation recording**: In the real world, volunteers manipulate the garments from a given initial pose, and the point cloud sequences of garment movements and the hand poses are recorded with a multi-view RGB-D camera setup. (3) **Coarse alignment**: We replay the manipulation process with scanned garment models and recorded hand poses in a simulator [15] with physics-based animation. In this way, the garment poses can be coarsely aligned with real-world observations. (4) **Fine pose fitting**: Finally, we use an optimization-based approach to fit the coarsely aligned shape onto the corresponding point cloud sequence from the real world. (5) **Verification**: After the process, we verify the annotation by checking the accuracy at predefined key points, which is called *Grid Layout*, and visually inspect whether the fitted pose is satisfactory. The GarmentTwin pipeline can effectively fit the garment pose with good accuracy ($\sim 0.2cm$) and without penetration.

With GarmentTwin, we can collect real-world garment pose flow during manipulation, which enables us to construct a large-scale real-world garment pose dataset, **ClothPose**. ClothPose is built upon a real-world garment repository, **Garment3D**, which includes 600 garments of 10 categories. We ask volunteers to manipulate the garments with predefined operations, *i.e. folding* and *randomization*. As a result, we have 30K point cloud sequences with annotated poses for every frame.

Although we annotate the complete pose and are primarily interested in non-rigid pose estimation tasks (NRPE), the annotation can also support incomplete tasks such as non-rigid reconstruction (NRR). Therefore, we benchmark NRR with different baseline approaches to attract a broader audience. Specifically, we adopt DynamicFusion [34] and DeepDeform [6] for NRR task, and GarmentNets [9] for NRPE task.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to accomplish garment pose recording with complex pose configurations in natural manipulation tasks in the real world. To achieve this, we propose a novel recording pipeline **GarmentTwin**.
- With GarmentTwin, we propose a large-scale real-world dataset, **ClothPose**, which includes an asset dataset, Garment3D, with 600 garments of 10 categories, and a task dataset with RGB-D sequences of garments in manipulation. This dataset allows researchers to conduct garment pose research directly in the real world for the first time.
- We benchmark two relevant non-rigid tasks namely reconstruction and pose estimation on our ClothPose dataset with different baselines. We are especially interested in facilitating the research in non-rigid pose estimation tasks, which are less explored as no proper benchmark exists before.

2. Related Works

2.1. Sensors for Localization

Researchers have developed many localization technologies to capture an object's location in the real world. However, garment pose estimation presents unique requirements to the sensors, including:

· **Sub-centimeter-level accuracy** within a space of $1.5m^3$, which is a typical workspace for folding.

· **Robustness to the occlusion.**

· **Lightweight** pose indicators (*i.e.* the signal receiver or the marker) attached to the garment.

· **Flexibility** to adapt to the garment deformation.

To the best of our knowledge, current localization technology cannot satisfy all these requirements. The accuracy requirement alone rules out most of the active sensors. For example, IMU [14] can provide quite accurate angular information, but has poor location accuracy, which can drift away $9m$ in 20 seconds even after carefully calibration [14]. GPS [16] has decimeter-level accuracy, and cannot be used indoors or for 3D localization. UWB [61] has an accuracy of up to $5cm$ in indoor environments, but occlusion can easily influence it. WiFi [23] has only decimeter-level accuracy. Magnetic sensors' accuracy worsens as the receiver gets farther from the magnetic source, with an empirical accuracy of $2.6cm$ at $29cm$ [7]. Ultrasonic waves [22] cannot bypass occlusion. Bragg grating [41] is too expensive and has limited bending ability. While multi-sensor fusion may improve accuracy [8], the receiver's larger size makes it hard for sensors to attach to garments. Additionally, we are not aware of a multi-sensor fusion solution that can achieve 3D pose estimation at the sub-centimeter level. Given the current status of localization sensors, we turn to indirect solutions to record garment poses.

2.2. Garment Dataset

The current 3D garment dataset can be roughly categorized as either *scanned* or *simulated*. Simulated datasets

Dataset	# Garments	# Category	# Frame	Deformation	Dynamics	Annotation
Synthetic						
TailorNet[39]	20	/	*	*	simulated	*
Cloth3D[4]	11.3K	4	*	*	simulated	*
Real						
VideoFolding[46]	-	2	304K	large	real	arm keypoints
DeepFashion3D[62]	563	10	2K	mild	static	feature lines
DeepDeform[6]	-	1	∼6K	mild	real	sparse keypoints
ClothPose	600	10	30K	large	real	per-vertex pose

Table 1. Comparison with other garment-related datasets. (1) "# Garments" column compares the number of garment models provided in the dataset. VideoFolding and DeepDeform are both 2.5D based, they do not provide complete garment models. (2) "Deformation" means if the deformations in the dataset are complex. "mild" means the operation on garments will cause gentle deformation. The operations are usually hanging or swinging. (3) In the "Dynamics" column, "static" means not dynamics, "real" means the dynamics follow real physics, while "simulated" not. (4) * means it can be synthesized.

[4, 39, 40] often have larger dataset sizes but lack the realistic texture and plausible dynamics since existing simulation techniques cannot fully replicate the garment motion dynamics [58]. Therefore, we favor the real-world scanning path and are particularly interested in *scanned* garment dataset. The BUFF dataset [60] provides very few scans and sequences, while 3DPW [47] focuses more on human body pose, though it also provides 18 clothes models. These on-body datasets cannot separate the garment pose from the body pose. Later, Deep Fashion3D [62] proposes a garment model repository that contains 563 garments. Since they do not provide high-resolution mesh models, and target on-body reconstruction tasks, the pose varieties are rather limited. Thus, they cannot be used to support garment pose estimation and tracking tasks. From the task perspective, a video dataset proposed by [46] is more relevant. However, they only have a limited number of garments, and the annotations are only key points of the operator's arms and hands. A detailed comparison can be referred to in Table 1.

2.3. Non-rigid Reconstruction

The study of reconstructing general non-rigid objects with large deformation has a long history [48]. However, early works [13] had slow offline runtimes and required slow and simple motion. A milestone work is DynamicFusion [34], which is the first approach to achieve non-rigid reconstruction in real-time. Since DynamicFusion, many following works have been proposed [21, 42, 43] that improve tracking quality [21], enable the topology changes [42, 43], or consider additional information like geometry, albedo, and motion [17]. Deepdeform [6] proposed a new learning-based sparse-to-dense reconstruction approach that showed good performance on the accompanying dataset. Recently, Xu et al. [51] integrates both tactile and visual perception for non-rigid object reconstruction. While the typical non-rigid reconstruction (NRR) task aims to reconstruct only the visible surface, there is also another line of works targeting 4D holistic object reconstruction [12, 11, 26, 35, 36, 37, 27]. These methods can reconstruct deformable objects, usually the human body, with complete mesh. However, since they do not seek to fit the mesh into a configuration space. Thus, they cannot obtain the object pose.

Non-rigid Pose Estimation Rigid pose estimation has been widely studied in the computer vision community [28, 29, 52]. In contrast, non-rigid pose estimation can be regarded as a form of template-driven reconstruction, whether the template can be either explicit [30] or implicit [9]. This area has been extensively researched in the human body modeling community. However, although the human body is non-rigid, its deformation is typically mild, and it can be considered an articulated object [10, 54, 56, 31, 55]. More recently, on-body clothes reconstruction [40, 60, 32, 45, 57, 59, 5, 19] has received increasing attention. However, due to the constraints of the human body, on-body clothes cannot have highly non-rigid deformations. Yang *et al*. [57] propose to reconstruct the garment model from a single image with the aid of human body pose estimation along with the garment parameters. Yu *et al*. [59] also propose a simulate-and-fit pipeline like ours, however since the garment deformation is caused by body motion, only cloth-body collision is considered in the pipeline. Such strategy cannot be applied to off-body garment manipulation, where cloth-cloth collision is much more severe. Besides, the mass-spring system used in [59] cannot handle the cloth-cloth collision during physics-based simulation without additional constraints, such as [50], which was not considered in [59]. Recently, Xue et al. [53] proposed a simulated system for non-rigid garment manipulation data collection. It also presents a 4D non-rigid pose estimation method.

Compared to the NRR task, the NRPE task presents unique challenges, as it involves handling the unseen parts of the surface and seeks the correspondence in configuration space. GarmentNets [9] was recently proposed to estimate garment pose from the in-grasping configuration. In this work, we benchmark it using our proposed ClothPose

on different manipulation tasks (Sec. 4).

3. GarmentTwin

3.1. Garment3D: A Large-scale Real-World Garment Repository

We purchased 600 real-world garments. The following will describe how to categorize them, and scan them into 3D models.

Repository Statistics. We follow the categorization as in [62], and also have 10 categories. We report the basic statistics of the asset dataset in Table 2. The data samples are presented in the supplementary materials.

Category	# Object	# Vertex	# Triangle
Long-sleeve coat	171	408K	814K
Short-sleeve coat	75	317K	632K
None-sleeve coat	10	171K	340K
Long trousers	49	381K	759K
Short trousers	35	166K	330K
Long-sleeve dress	52	607K	1210K
Short-sleeve dress	46	565K	1127K
None-sleeve dress	31	390K	776K
Long skirt	55	415K	827K
Short skirt	76	189K	375K

Table 2. The statistics of each garment category. "# Vertex" and "# Triangle" are the average numbers of vertices and triangles in the category respectively.

Scanning. To scan each garment, we place it on a mannequin until it reaches static equilibrium. The pose in this state is defined as the *rest pose*. We then use an Einscan pro 2x 2020 3D scanner to scan the clothes model. As shown in Fig. 2, to reduce post-processing procedures, the mannequin is made of reflective black material, which is hard to be scanned by the structural light-based scanner, so that the scanning result will be the garment alone. The scanning process typically takes between 20 and 60 minutes, depending on the garment size. After scanning, we use surface reconstruction, hole filling, and texture fusion to fix the models' surfaces and textures.

Grid Layout. We define a layout of sparse keypoint on each garment for each category, called *Grid Layout*. The layout serves two purposes: to aid in the scanning process, particularly for textureless garments, and to serve as a verification checkpoint for pose annotation accuracy. We give an example in Fig. 2. The definition of the grid layout for each category can be referred to in the supplementary materials.

3.2. Real-World Recording Setup

After acquiring the garments, our next step is to manipulate and record them in the real world. The entire recording setup is illustrated in Fig. 1.

Hardware Setup. We use four Azure Kinect cameras to capture RGB-D sequences of the garment manipulation process. Each camera streams images at a sampling rate of 30 FPS, with a resolution of 640×480 for both RGB and depth streams. These cameras are setup with hardware synchronization.

Calibration. To calibrate multi-cameras, we design a multi-facet checkerboard. We register the point cloud from the other cameras to a reference camera using [38], and then use the transformed pose as initialization to calibrate the multi-camera with the Multical library [3]. Using this approach, we achieve a calibration error of $\sim 5mm$. After calibration, we fuse point clouds from each depth camera into an integral scene to obtain an *almost* complete outer observation of the garment state.

Manipulation. We conduct two different manipulation tasks on garments, namely *folding* and *randomization*. They both start from the same *initialization* process.

Initialization: For both manipulation tasks, we assume that the garments are initially laid on a flat surface in a simple configuration, as shown in Fig. 1. It is not a strict pose requirement, as long as the garment is flattened and aligned in a similar direction within the same category. We give examples in the supplementary materials.

Folding: In this task, the operator is asked to complete the garment folding process. Though different people may have different preferences when folding clothes, we ask the operator to follow specific procedures. For example, folding a long-sleeve coat takes three steps: (1) The operator grasps the left cuff and places it on the waistline. (2) He/She grasps the right cuff and places it on the waistline. (3) He/She grasps the neckline with two hands and places it on the waistline. For the procedures of the remaining categories, please refer to the supplementary materials.

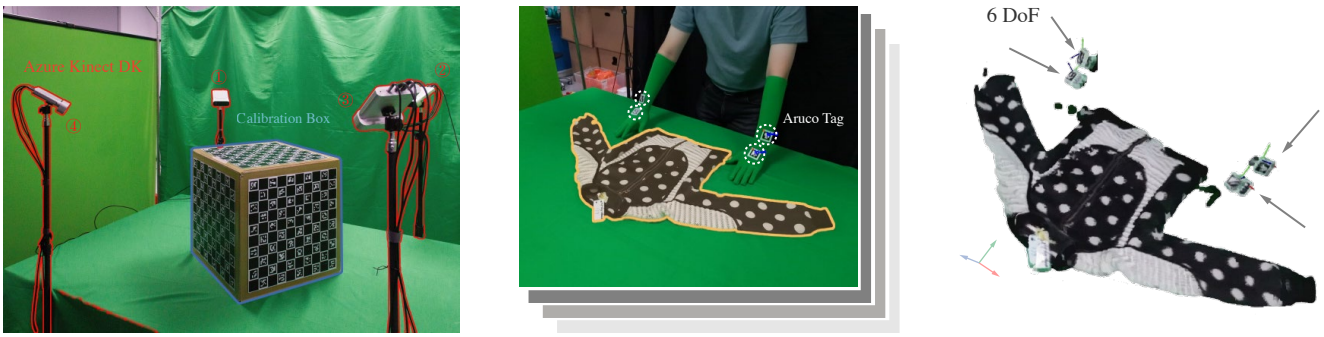
Randomization: In this task, the operator can manipulate the garment into any configuration they want with one move, such as picking-and-placing, pushing. The resulting garment configurations will be less regular than those in the folding process.

Recording. During manipulation, we cover the table and the surroundings with green sheets to facilitate the segmentation of the garment being manipulated from the background. We also attach Aruco markers on both hands of the operator to obtain the hand wrist pose by solving a PnP problem [24]. We then project the multi-view RGB-D images to an RGB-D point cloud scene, with the foreground filtered out, and annotate when the hands are performing grasping and releasing.

3.3. Coarse Alignment in Simulation

After obtaining the cleaned scene point clouds and hand poses, we use the scene point clouds as a reference and re-

https://www.einscan.com/handheld-3d-scanner/einscan-pro-2x/



(a) (b) (c)

Figure 1. **(a)**. Hardware recording setup. 1-4 are Azure Kinect cameras. A table covered with a green sheet is sited in the middle, and the multi-facet checkerboard box is placed on the table. **(b-c)** Segmentation of foreground garment, and multi-camera fusion for scene point cloud.

Figure 2. **Left**: The scanning setup includes the mannequin and the scanner. **Right**: The top is the scanned rest pose model and the bottom is the grid layout. The marker is scaled up for display, the actual size can be viewed in Left.

play the hand poses to manipulate the corresponding garment model in the simulator to align it. The process is illustrated in Fig. 3.

Physics-based Animation. We first simulate the garment dynamics with XPBD-based techniques [44] in RFUniverse [15] to deform the garment model. Then, we load the scanned garment model into RFUniverse along with the clean scene point clouds. We adjust the global orientation to align with the scene point clouds which is one-time effort for a sequence. Next, we place virtual hands in the scene based on the recorded poses. When the hand grasps the garment, it pins garment vertices within a distance of $1cm$, causing them to move along with the hand, and drive the garment deformation accordingly. To note, since we only annotate the hand wrist pose, the fingers will not move. The hand model loaded is only used for calculating the nearest vertices. When the hand releases the garment, the attachment between the garment model and hand model is removed, causing the vertices to fall down due to gravity.

Although the cloth simulation is not perfect, it replicates the essence of the process, which is sufficient for our final refinement optimization.

3.4. Fine Pose Fitting

We define the garment mesh model in rest pose as $\mathcal{M}^{rest} = (V^{rest}, E^{rest})$, where V^{rest} represents the vertex locations and E^{rest} the edges between vertices. The coarsely aligned garment mesh models are $\mathcal{M}_t^{coarse} = (V_t^{coarse}, E_t^{coarse})$, t means the frame index and starts from 0. To note, E_t^{coarse} will not change throughout the manipulation process, and it is also the same with E^{rest}, thus they naturally correspond to each other. The fused point clouds from depth cameras are denoted as $\mathcal{P}_t$.

Since after calibration, the depth camera errors are within $1cm$, we consider the fused point clouds $\mathcal{P}_t$ can be regarded as *almost complete* and *accurate* for outer observations. To fully utilize the point clouds, we adopt a three-stage strategy to refine the garment mesh $\mathcal{M}_t^{coarse}$ to the point cloud $\mathcal{P}_t$.

Stage 1: Non-rigid ICP. In the first stage, we first apply a non-rigid iterative closest point (ICP) algorithm from [1] between $\mathcal{M}_t^{coarse}$ and $\mathcal{P}_t$. This non-rigid ICP process tries to match the vertices in each $\mathcal{M}_t^{coarse}$ to the corresponding $\mathcal{P}_t$, and to deform the vertices to the matched points as close as possible. However, the result from this algorithm may contain severe self-intersections and large deformations. We denote the mesh model result from this step as $\mathcal{M}_t^{s1}$.

Stage 2: Resolving Interpenetration & Natural Deformation. To eliminate these artifacts, we add two terms to

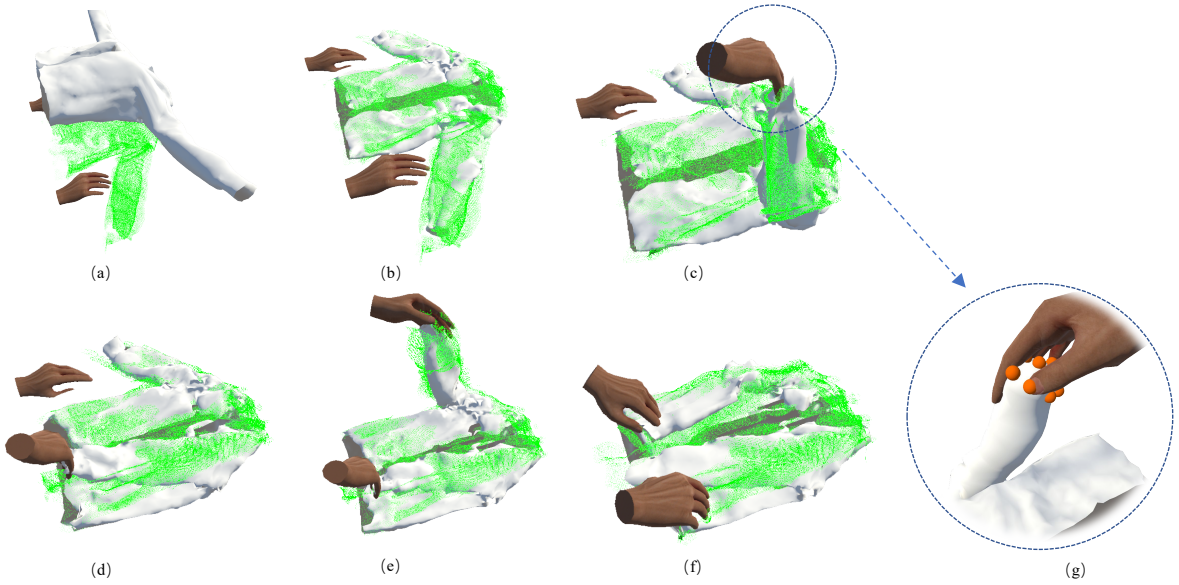


Figure 3. Coarse Annotation in Simulation. **(a)**. We load the garment mesh model (color/texture are removed to make the alignment clearer), and scene point clouds (in Green) into RFUniverse. **(b)**. We adjust the global orientation of the garment to align with the point cloud from initialization stage. **(c)**. When a grasping state is activated for a hand, it will pin the nearest points on the garment to the hand **(g)** to drive the deformation. **(d)-(f)** Continuing a folding process according to the hand wrist poses and states.

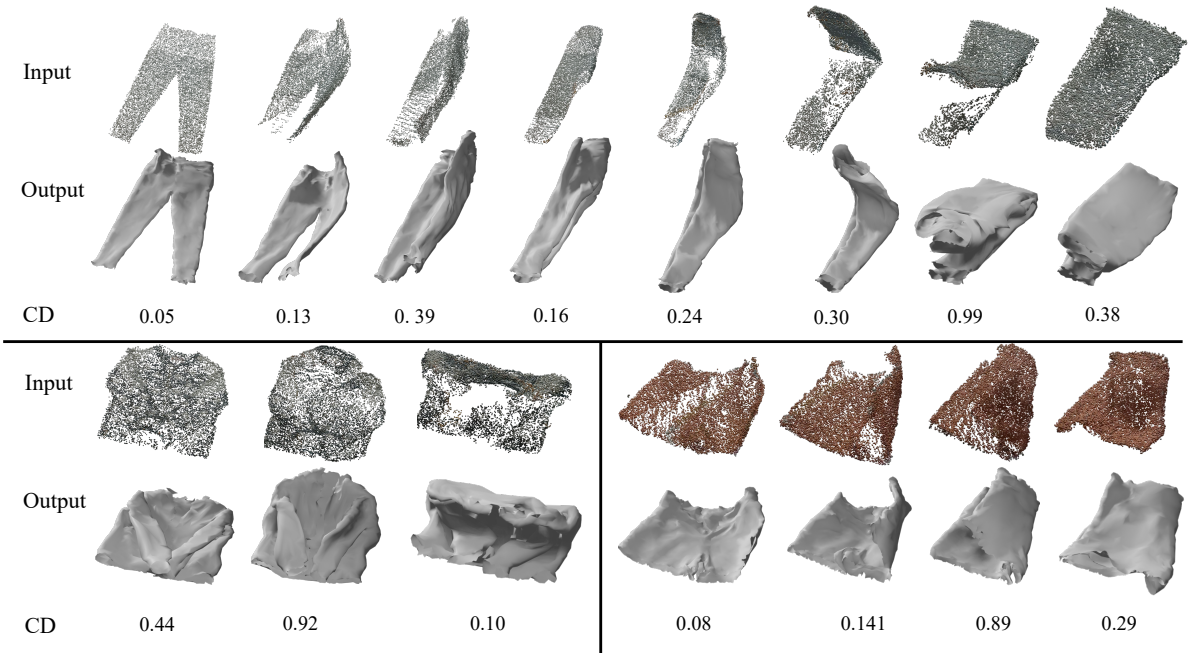


Figure 4. Qualitative results of garment pose annotation. **Input**: input point clouds. To note, some point clouds may seem posit in a simple configuration (*e.g.* the fourth trousers in the first row), they may have already folded once. **Output**: Fitted results after GarmentTwin pipeline. **CD**: The Chamfer distance between the output mesh and the input point cloud.

control the interpenetration and deformation. Denote the result mesh in this stage as $\mathcal{M}_t^{s2}$, an energy function E_{s2} is defined as follow:

$$\begin{aligned} E_{s2}(\mathcal{M}_t^{s2}, \mathcal{M}_t^{s1}, \mathcal{M}^{rest}) =& E_{L_2}(\mathcal{M}_t^{s2}, \mathcal{M}_t^{s1}) \\ &+ \lambda_1 E_{pene}(\mathcal{M}_t^{s2}) \\ &+ \lambda_2 E_{reg}(\mathcal{M}_t^{s2}, \mathcal{M}^{rest}). \end{aligned} \tag{1}$$

E_{s2} is composed of three terms. E_{L_2} is the L_2 loss between the result mesh $\mathcal{M}_t^{s2}$ and the result of the non-rigid ICP algorithm [1] $\mathcal{M}_t^{s1}$. E_{pene} is a barrier function term borrowed from [25] to prevent self-intersections by penalizing primitive pairs (*i.e.* vertex-vertex, vertex-edge pairs) of which the distances are below a certain threshold $\hat{d} = 0.0005$. Additionally, as in [25], floating-point continuous collision detection (CCD) is also applied to avoid self-intersections by clamping the step size. It can dynamically detect the potential primitive pairs to collide between optimization steps. Finally, E_{reg} is an as-rigid-as-possible (ARAP) [20] regularization term, it can regulate the deformation to a natural state. We optimize this objective function with the Projective Newton method. The result is now intersection-free without large deformation. However, it still does not perfectly fit the point cloud. To solve this problem, stage 3 is introduced.

Stage 3: Fitting To Point Cloud. To perfectly fit the point cloud, we add a Chamfer Distance term to the optimization function. Denote the result mesh in this stage as $\mathcal{M}_t^{s3}$, the optimization function E_{s3} is as follow:

$$\begin{aligned} E_{s3}(\mathcal{M}_t^{s3}, \mathcal{P}_t, \mathcal{M}^{rest}) =& \lambda_3 E_{champf}(\mathcal{M}_t^{s2}, \mathcal{P}_t) \\ &+ \lambda_4 E_{pene}(\mathcal{M}_t^{s3}) \\ &+ \lambda_5 E_{reg}(\mathcal{M}_t^{s3}, \mathcal{M}^{rest}). \end{aligned} \tag{2}$$

In E_{s3}, E_{champf} is the Chamfer Distance [2] between the mesh to be optimized $\mathcal{M}_t^{s2}$ and the point cloud $\mathcal{P}_t$, E_{reg} and E_{pene} are the same as in E_{s2}. We also optimize this function with the Projective Newton method. Since the result of non-convex optimization is sensitive to the initialization, we choose $\mathcal{M}_{t-1}^{s3}$ to be the initial value of stage 2, $\mathcal{M}_t^{s2}$ to be the initial value of stage 3. Specifically, $\mathcal{M}_{-1}^{s3}$ here is defined as $\mathcal{M}^{rest}$ if frame index t starts from 0.

Implementation Details. In implementation, we set $\lambda_1 = 1e7$, $\lambda_2 = 1$, $\lambda_3 = 3$, $\lambda_4 = 1e7$, $\lambda_5 = 1$. The Projective Newton optimization for both stages will iterate up to 100 times, or it can end earlier if the L_2 norm of the search direction is less than 0.0005. The optimization can be done in 5 seconds per frame.

3.5. Annotation Verification

Admittedly, though we can fit the accurate point cloud with a reasonably low energy, we still cannot guarantee pose in the occluded part is right, as they cannot be observed in the point clouds. We still need to verify the optimization results. To verify the optimization results, we conduct it in both quantitative and qualitative ways: (1) For the visible part, we check the errors at the predefined layout points. The average corresponding distance is $2.3mm$. (2) For the invisible part, we ask the annotator to check if the fit between the garment mesh and the point cloud is visually pleasant. To note, with the E_{champf}, we can also manually select the corresponding pairs before optimization, which can further improve the accuracy of garment fitting. We illustrate samples after optimization in Fig. 4.

3.6. ClothPose Dataset Statistics

Since the depth cameras can work at 30 FPS, the volunteers can manipulate the garment at a natural speed. They usually finish the folding task in 5 seconds and the randomization task in 2 seconds. After cleaning, we filter out the consecutive frames with minor deformation, which happens when the volunteers change the operation (*e.g.* in the duration of after placing the left cuff and transiting to grasp the right cuff).

We manipulated each garment once for the *folding* task and collected 600 videos in total, which contain 18K frames. For the randomization task, we randomly manipulated the garment three times, resulting in 1800 video clips in total, which contain 12K frames. We select $1/5$ garments (item number is rounded up to the nearest integer) from each category as unseen and the corresponding frames as the test set for evaluating the learning algorithms.

4. Experiments

Though we are mostly interested in garment pose estimation tasks, to broader usage of this dataset, we also benchmark the dataset with non-rigid reconstruction tasks. Additionally, we conduct ablative study on the simulation-fitting process. More discussion on how to process the scanned garment model, physics engine choices for simulation can be referred to supplementary materials.

4.1. Metrics

Metrics for Non-rigid Reconstruction, NRR task. We follow the metrics in [6] and report the deformation error (Def-err) and geometry error (Geo-err) in centimeters.

Metrics for Non-rigid Pose Estimation, NRPE task. We put it in a category-level setting, as currently, the only baseline approach GarmentNets [9] does. We adopt two metrics, namely Chamfer Distance(D_{champf}) and Correspondence Distance (D_{corr}). D_{champf} calculates Chamfer distance in centimeters between the reconstructed mesh points and the ground-truth mesh points. This metric can evaluate the quality of surface reconstruction. D_{corr} calculates the L2 distance in centimeters between the reconstructed mesh vertices and the ground-truth mesh vertices. To note, since D_{corr} requires one-to-one mapping between the vertices (*i.e.* configuration space), for GarmentNets, we calculate this based on the NOCS [49] coordinates (*i.e.* each point on the predicted mesh will find the closest point on the ground-truth mesh in NOCS).

4.2. Results

Results on Non-rigid Reconstruction. We adopt DynamicFusion [34] and DeepDeform [6] as the baseline for the NRR task. As shown in Table 3.

Method	Def-err ↓	Geo-err ↓
DynamicFusion [34]	8.43	1.37
DeepDeform [6]	4.89	0.66

Table 3. Results on NRR task with the baseline of DynamicFusion and DeepDeform. Such methods are category-agnostic, thus we calculate the accuracy from the whole dataset. ↓ means the lower the better.

Results on Non-rigid Pose Estimation. Finally, we adopt GarmentNets [9] as the baseline for the category-level NRPE tasks on garments, as it is currently the only learning-based approach for category-level pose estimation. As shown in Table 4, and demonstrated in Fig. 5, we can observe failure cases when taking GarmentNets for pose estimation. For example, the overall translation and scale errors, the structure incompleteness, and the correspondence errors.

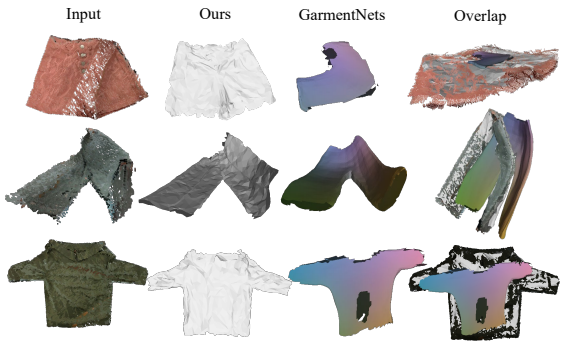


Figure 5. **Input**: input point clouds. **Ours**: Fitted results after GarmentTwin pipeline. **GarmentNets**: Predicted results of GarmentNets. **Overlap**: We put all these into one scene so that we can check the quality of the fitting and GarmentNets estimation. The viewpoint might be different. More results with GarmentNets are given in supplementary materials, and the overlap column is better viewed in 3D mode, which will also presented in the supplementary video.

The reason why GarmentNets fail to achieve good performance on our dataset is due to its design for reconstructing garment pose from static observations. In the original work [9], the dataset used to train GarmentNets consists of similar poses, where the garments are grasped by one hand and hanging in the air. However, our dataset focuses on the garment poses with manipulation dynamics. Data samples for a specific configuration such as "being grasped and hanging in the air" occupy only a small portion of the overall manipulation sequence. Thus, ClothPose have a much wider variety of poses than GarmentNets dataset (we will provide a visual comparison in the supplementary material). In such scenarios, we have observed that GarmentNets cannot effectively capture the pose changes. Thus, we believe that non-rigid pose estimation for garments is a more challenging and less explored area of research. We hope that our benchmark will encourage more research in this direction.

Category	D_{chamf} ↓	D_{corr} ↓
Long-sleeve coat	1.54	5.27
Short-sleeve coat	1.62	6.44
None-sleeve coat	1.58	5.69
Long trousers	1.27	4.38
Short trousers	1.42	4.59
Long-sleeve dress	1.66	5.73
Short-sleeve dress	1.63	5.44
None-sleeve dress	1.75	6.21
Long skirt	1.94	6.99
Short skirt	1.82	6.91

Table 4. Results on NRPE of GarmentNets on each category of ClothPose. ↓ means the lower the better.

4.3. Ablative Study on Simulation-Fitting Process

We carry out the ablative study on simulation-fitting process with a long trousers folding sequence. As shown in Fig. 6, without mesh produced by the coarse alignment process ("w/o coarse"), directly optimizing from rest pose can make the Chamfer distance fail to match the points between garment mesh and point cloud, and thus the shape can be very different from the observation. Without ARAP term ("w/o reg term"), it can make the Hessian matrix in penetration term fails to keep semi-definite. For this part, readers can refer to IPC paper [25]. Thus, it cannot finish the sequence due to the numeric error for later frames. Without penetration term ("w/o penetration term"), it can easily penetrate among layers. The dark area is the penetrated area. The Chamfer distances for the "w/o coarse", "w/o reg term", "w/o penetration" and "full" is 1.26, N/A (since it encounters a numeric error in the end), 0.48, 0.23 and 0.16 respectively.

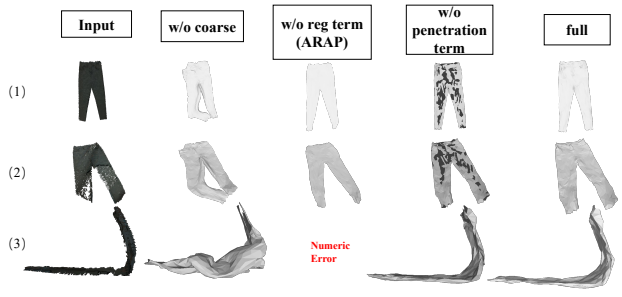


Figure 6. Ablative study on pose fitting process. (1) and (2) are frame #1 and #5, (3) is frame #25 in the selected sequence.

4.4. Limitation

We think there are two major limitations: 1. It requires a garment mesh model, which requires a whole scanning setup, and it is troublesome. 2. The pose alignment and fitting process takes a few seconds. Ideally, if it can run in real-time, the method can be used for wider scenarios. Such limitations are also the improvement directions we are particularly interested in.

5. Conclusion and Future Works

In this work, we propose a novel pipeline, GarmentTwin to record real-world data of garment pose based on its dynamic movements. With the recording pipeline GarmentTwin, we take the first step towards constructing a real-world large-scale dataset for garment pose estimation. Using the dataset ClothPose, we benchmark two relevant tasks with multiple baselines. We hope the proposal of this benchmark can encourage more attention to this challenging and exciting direction. Besides, since our recorded garments are in manipulation, we are interested in applying them to robotic applications.

Acknowledge This work was supported by the National Key R&D Program of China (No. 2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and Shanghai Science and Technology Commission (21511101200).

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 5, 7

[2] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1977. 7

[3] Oliver Batchelor. multical. https://github.com/oliver-batchelor/multical, 2022. 4

[4] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 1, 3

[5] 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

[6] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020. 1, 2, 3, 7, 8

[7] Dongyao Chen, Mingke Wang, Chenxi He, Qing Luo, Yasha Iravantchi, Alanson Sample, Kang G. Shin, and Xinbing Wang. Magx: Wearable, untethered hands tracking with passive magnets. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, MobiCom '21, pages 269–282, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2

[8] Zhenghua Chen, Han Zou, Hao Jiang, Qingchang Zhu, Yeng Chai Soh, and Lihua Xie. Fusion of wifi, smartphone sensors and landmarks using the kalman filter for indoor localization. *Sensors*, 15(1):715–732, 2015. 2

[9] Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. *arXiv preprint arXiv:2104.05177*, 2021. 1, 2, 3, 7, 8

[10] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 3

[11] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):1–16, 2017. 3

[12] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 3

[13] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi. 3d scanning deformable objects with a single rgbd sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–501, 2015. 3

[14] Jeff Ferguson. Calibration of deterministic imu errors. 2015. 2

[15] Haoyuan Fu, Wenqiang Xu, Ruolin Ye, Han Xue, Zhenjun Yu, Tutian Tang, Yutong Li, Wenxin Du, Jieyi Zhang, and Cewu Lu. Demonstrating rfuniverse: A multiphysics simulation platform for embodied ai. 2, 5

[16] Mohinder S Grewal, Lawrence R Weill, and Angus P Andrews. *Global positioning systems, inertial navigation, and integration*. John Wiley & Sons, 2007. 2

[17] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 3

[18] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 1

[19] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud se-

quences. *Advances in Neural Information Processing Systems*, 34:27940–27951, 2021. 3

[20] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM transactions on Graphics (TOG)*, 24(3):1134–1141, 2005. 7

[21] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016. 3

[22] AR Jiménez and F Seco. Ultrasonic localization methods for accurate positioning. *Instituto de Automatica Industrial, Madrid*, 2005. 2

[23] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pages 269–282, 2015. 2

[24] Vincent Lepetit, Francesc Moreno-Noguer, and P Fua. Epnp: Efficient perspective-n-point camera pose estimation. *Int. J. Comput. Vis*, 81(2):155–166, 2009. 4

[25] Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy R Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M Kaufman. Incremental potential contact: intersection-and inversion-free, large-deformation dynamics. *ACM Trans. Graph.*, 39(4):49, 2020. 7, 8

[26] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12706–12716, 2021. 3

[27] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2022. 3

[28] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 3

[29] Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31:1072–1083, 2022. 3

[30] 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. 3

[31] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. Handtailor: Towards high-precision monocular 3d hand recovery. *arXiv preprint arXiv:2102.09244*, 2021. 3

[32] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16082–16093, 2021. 3

[33] Roberto Martín-Martín, Clemens Eppner, and Oliver Brock. The rbo dataset of articulated objects and interactions. *The International Journal of Robotics Research*, 38(9):1013–1019, 2019. 1

[34] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2, 3, 8

[35] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019. 3

[36] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 3

[37] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12851–12860, 2022. 3

[38] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017. 4

[39] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 3

[40] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2242–2251, 2019. 3

[41] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement science and technology*, 8(4):355, 1997. 2

[42] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017. 3

[43] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2646–2655, 2018. 3

[44] Virtual Method Studio. 5

[45] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 1–18. Springer, 2020. 3

[46] Andreas Verleysen, Matthijs Biondina, and Francis Wyffels. Video dataset of human demonstrations of folding clothing

for robotic folding. *The International Journal of Robotics Research*, 39(9):1031–1036, 2020. 1, 3

[47] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 3

[48] Michael Wand, Bart Adams, Maksim Ovsjanikov, Alexander Berner, Martin Bokeloh, Philipp Jenke, Leonidas Guibas, Hans-Peter Seidel, and Andreas Schilling. Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data. *ACM Transactions on Graphics (TOG)*, 28(2):1–15, 2009. 3

[49] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 7

[50] Longhua Wu, Botao Wu, Yin Yang, and Huamin Wang. A safe and fast repulsion method for gpu-based cloth self collisions. *ACM Transactions on Graphics (TOG)*, 40(1):1–18, 2020. 3

[51] Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, and Cewu Lu. Visual-tactile sensing for in-hand object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8803–8812, 2023. 3

[52] Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval. *arXiv preprint arXiv:2112.07334*, 2021. 3

[53] Han Xue, Wenqiang Xu, Jieyi Zhang, Tutian Tang, Yutong Li, Wenxin Du, Ruolin Ye, and Cewu Lu. Garmenttracking: Category-level garment pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21233–21242, 2023. 3

[54] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020. 3

[55] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2750–2760, 2022. 3

[56] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 3

[57] Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG)*, 37(5):1–14, 2018. 3

[58] Hang Yin, Anastasia Varava, and Danica Kragic. Modeling, learning, perception, and control methods for deformable object manipulation. *Science Robotics*, 6(54), 2021. 3

[59] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[60] 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. 3

[61] Yuan Zhou, Choi Look Law, and Jingjing Xia. Ultra low-power uwb-rfid system for precise location-aware applications. In *2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 154–158. IEEE, 2012. 2

[62] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision*, pages 512–530. Springer, 2020. 3, 4