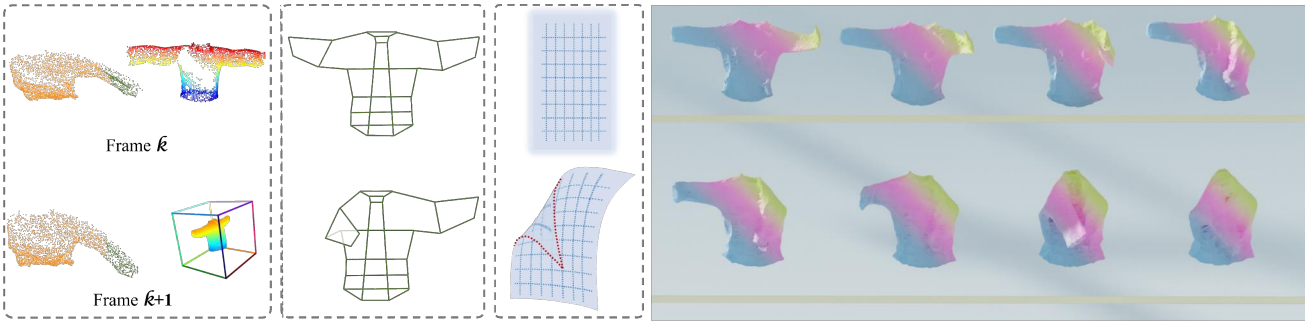
GaPT-DAR: Category-level Garments Pose Tracking via Integrated 2D Deformation and 3D Reconstruction

Li Zhang [1,2,3], Mingliang Xu*[1], Jianan Wang [3], Qiaojun Yu [4], Lixin Yang [4], Yonglu Li [4], Cewu Lu [4], Rujing Wang [2†], Liu Liu [5†]
[1] University of Science and Technology of China, Hefei, China
[2] Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[3] Astribot, Shenzhen, China. [4] Shanghai Jiao Tong University, Shanghai, China
[5] Hefei University of Technology, Hefei, China
zanly20@mail.ustc.edu.cn



(*a*) Inputs (*b*) 2D Projection (*c*) Warping Vector (*d*) Outputs: tracking results in task space

Figure 1. Category-level **Ga**rments **P**ose **T**racking framework with integrated 2D **D**eformation **A**nd 3D **R**econstruction (**GaPT-DAR**). Given the partial point clouds from adjacent frames and pose prediction result from the previous frame, we propose a garment pose tracking pipeline consisting of 3D-2D projection, 2D deformation learning, and 2D-3D reconstruction steps. The output is the tracked pose (complete point cloud) of garments in task space.

Abstract

*Garments are common in daily life and are important for embodied intelligence community. Current category-level garments pose tracking works focus on predicting point-wise canonical correspondence and learning a shape deformation in point cloud sequences. In this paper, motivated by the 2D warping space and shape prior, we propose **GaPT-DAR**, a novel category-level **Ga**rments **P**ose **T**racking framework with integrated 2D **D**eformation **A**nd 3D **R**econstruction function, which fully utilize 3D-2D projection and 2D-3D reconstruction to transform the 3D point-wise learning into 2D warping deformation learning. Specifically, GaPT-DAR firstly builds a Voting-based Project module that learns the optimal 3D-2D projection plane for maintaining the maximum orthogonal entropy during point projection. Next, a Garments Deformation module is designed in 2D space to explicitly model the garments warping procedure with deformation parameters. Finally, we build a Depth Reconstruction module to recover the 2D images into 3D warp field. We provide extensive experiments on VR-Folding dataset to evaluate our GaPT-DAR and the results show obvious improvements on most of the metrics compared to state-of-the-arts (*i.e. *GarmentNets [8] and GarmentTracking [32]). More details are available at https://sites.google.com/view/gapt-dar.*

1. Introduction

Garments, as a type of non-rigid object, are pervasive in daily life. Visual pose tracking for garments is crucial for the embodied intelligence community and beneficial for many downstream tasks such as interactive perception [36], object manipulation [31], imitation learning [38],

*Li Zhang and Mingliang Xu contribute equally to this work. This work was done when Li Zhang was an intern at Astribot.
†Joint Corresponding Authors.

human-machine collaboration [16]. Different from rigid objects [41] and articulated objects [39] that can be regarded as the combination of finite rigid parts, garments poses are defined as a canonical point-wise learning and shape reconstruction task [8, 32] (*e.g*. NOCS coordinates). The intricate kinematic constraints inherent in garments bestow them with a significantly higher degree of freedom, rendering garment pose tracking, particularly for off-body garments, a challenging task.

Under this circumstance, recent works formulate the category-level garments pose tracking task as a dense prediction and reconstruction problem in a normalized coordinate space that is shared among the instances from one category, and build the encoder-decoder deep learning paradigm to solve it. GarmentsNets [8] pioneered the definition of category-level pose estimation for garments, aiming to map the observed partial surface of garments to the normalized canonical space. Following this setting, GarmentTracking [32] provides a large-scale garments manipulation video dataset and accurately tracks the vertex-level poses by concatenating features from adjacent frames. Despite the success of these works, several issues exist:

- They predict per-vertex corresponding canonical coordinates solely within the 3D point cloud space, which limits the model's ability to explicitly perceive the twisting process of the garments and diminishes interpretability.
- Except for some corner cases such as prom dresses, catwalk wear and other specially designed clothes, most common garments have natural geometric symmetry (front and rear surfaces) [5, 11, 26], which has not been paid attention by previous tracking methods. Recent studies [1, 6, 20] have demonstrated the effectiveness of incorporating human posture as prior information.

To address these challenges, we argue that the shape deformation of garments might be better perceived in 2D space rather than 3D observed point cloud. This approach leverages the more efficient modeling capabilities of 2D neural networks when processing 2D data, such as point sets. Motivated by this, we propose an integrated 2D shape **D**eformation **A**nd 3D **R**econstruction framework for category-level **Ga**rments **P**ose **T**racking task, namely **GaPT-DAR**, utilizing the 3D-2D transformation and 2D-3D reconstruction pipeline to boost the pose perception performance. In our GaPT-DAR for tracking task, given the partial point clouds at adjacent frames along with the estimated canonical shape from the previous frame, we build a weight-shared ResUNet3D [9] for two-frame 3D feature fusion. Next, the **Voting-based Projection** module is designed to predict the projection plane by a voting-offset scheme that is utilized for 3D-2D projection of the point clouds. The projection plane aims to modulate and isolate the maximum orthogonal project information entropy, thus minimizing information loss during the 3D-2D projection process. After this, we propose a **Garments Deformation** module that learns the deformation function applied in 2D point sets and outputs pose-sensitive mappings to explicitly model the warping procedure from canonical state to task state. Finally, to recover the point sets into a 3D mesh Task (output), we build a **Depth Reconstruction** module that uses a simple neural network to regress the per-pixel depth in task space.

We evaluate our GaPT-DAR on category-level pose tracking tasks for garments on VR-Folding dataset [32], compared with state-of-the-art GarmentsNets [8] and GarmentTracking [32]. Our contributions can be summarized as follows:

- We present GaPT-DAR, an efficient and effective end-to-end framework tailored for category-level garment pose tracking. Central to our approach is the integration of a 3D-2D shape deformation mechanism, enabling the learning of garment pose warping within the 2D space.
- In our GaPT-DAR framework, we harness geometric symmetry as prior information to identify the optimal projection plane for acquiring the 2D garment point sets. This strategy aids in the learning of the garment deformation function via explicit parametric deformation modeling.
- Extensive experiments validate the superior performance of GaPT-DAR compared to state-of-the-art methods in category-level garment pose tracking tasks. Codes will be made publicly available.

2. Related Work

2.1. Non-Rigid Pose Estimation and Tracking

Rigid pose estimation and tracking has been widely studied in the computer vision [40] community that aims to predict 6D poses for the whole object [27, 28, 35] or each rigid part of articulation [17–19, 30]. Different from rigid or articulated objects, non-rigid pose estimation and tracking can be defined as a canonical shape completion and reconstruction task [8, 20]. A milestone work to achieve non-rigid reconstruction is proposed by Newcombe *et al*. [23] that exploits GPU solver to boost the reconstruction speed. The following works attempt to improve the tracking and reconstruction quality by geometry modeling [25], sparse representation [4] or motion understanding [10]. Recently, to accurately track and reconstruct non-rigid deformations, many works pay attention to the on-body clothes pose understanding tasks that leverage the human body as shape prior for better perception performance [2, 12, 21]. For example, Yang *et al*. [34] propose to reconstruct the garment model from a single image with the aid of human body pose estimation along with the non-rigid parameters. Yu *et al*. [37] also propose a simulate-and-fit pipeline like ours, however since the non-rigid deformation is caused by body motion,

only cloth body collision is considered in the pipeline.

Although the human body might offer a good reference for understanding clothes poses, it also contains large non-rigid deformations thus limiting the pose perception generalization. Therefore, our work focuses on garment pose estimation and tracking with higher complicated cases such as during manipulation or severe collision [32].

2.2. Off-body Garments Perception

To fully understand the non-rigid objects pose, *e.g*. clothes or garments, GarmentsNets [8] proposes a simulation dataset for category-level pose estimation based on CLOTH3D [1] asset, in which the garments are lifted by a robot with random pick points and forced to be stable with gravity. This setting and its corresponding pose estimation task take RGB-D images as input to recover the complete garment shapes in canonical space, but still suffer from self-occlusion missing. When moving to the garments manipulation case, GarmentTracking [32] records the garments videos by VR hardware to build a VR-Folding dataset and firstly presents a category-level garment pose tracking task. This dataset fully considers dynamic scenes including include complex human actions and garment configurations, which are more challenging.

However, this manner tends to result in unclear warping information easily while also neglecting geometric details. Shape deformation of Garments can be found in cloth simulation [26], Virtual try-on [43], Reconstruction [14], *etc*. These methods perform warping using explicit techniques, which are both efficient and effective. Building on these insights, our work advances the garment tracking task by extending and refining these techniques.

3. Notation and Problem Statement

Following GarmentTracking [32], we take the PC NOCS, mesh NOCS, and partial point cloud observations as input. As depicted in Fig. 2, ***to avoid ambiguity and better clarify the data flow, we use the symbol "$*$" to indicate the data and feature flow of PC NOCS and "$'$" to indicate those of mesh NOCS***. For example, $V_w^*, S_w^*, \mathcal{F}_w^*$ are used to illustrate the data flow for the PC NOCS branch, while we utilize $V_m', S_m', \mathcal{F}_m'$ to represent for the mesh NOCS branch.

To achieve robust category-level garment pose tracking, our core idea is modeling garment tracking via an integrated 3D-2D shape deformation and 3D depth reconstruction mechanism. Mathematically, we formulate the inputs and outputs in the garment tracking task as follows: given the self-occlusion point cloud P_k from k-th frame and P_{k+1} from $(k+1)$-th frame as **input**, we use PC NOCS P_k^* and a rest-state mesh NOCS P_{k+1}' serving as auxiliary information. Our target is to **output** the complete observation (mesh Task) for $(k+1)$-th frame in task space. Note that P_k, P_{k+1}, and P_k^* are partial observations, while P_{k+1}' is complete.

In our GaPT-DAR, we conduct a projection from 3D points to 2D point sets, which is achieved through a projection plane. Then we go on to perform 2D deformation with 2D point sets. we recover the depth information in the last stage. Specifically, firstly, we fuse 3D-CNN features from the k-th and $(k+1)$-th frame observation and output the $(k+1)$-th frame garment pose P_{k+1}^* in NOCS, facilitating downstream network utilization. Subsequently, we demonstrate how to project 3D points into 2D point sets. This procedure is conducted via the optimal projection plane for all the candidate points, we defined the optimal projection plane by a normal vector V and a pivot point. Afterward, given the features $\mathcal{F}_w^*$ and $\mathcal{F}_m^{'}$ from PC NOCS branch and mesh NOCS branch, respectively, we perform garment deformation guided by TPS transformation parameter θ. Finally, we reconstruct the point-wise depth and output the tracking result (mesh Task), which represents the $(k+1)$-th frame complete point cloud in task space.

4. Methodology

In this Section, the four interconnected modules of our GaPT-DAR are illustrated in detail, including Inter-frame Feature Fusion, Voting-based Projection, Garment Deformation, and Depth Reconstruction. In what follows, we describe in brief how these attempts are implemented and explain our innovations in detail accordingly.

4.1. Inter-frame Feature Fusion

Taking P_k and P_{k+1} as input, we handpick ResUNet3D [9] to sever as the backbone, which enjoys the merit of perceiving fine-grained local details well. Empirically, to generalize different garments regardless of intra-class variance (without precise 3D CAD models), we leverage NOCS [28] to complement the garment-aware feature representation. As depicted in Fig. 2, we use per-point NOCS coordinate prediction from the k-th frame P_k^* for positional embedding due to the clearer geometric and structural information.

To learn semantically meaningful connections between task space and NOCS, a fusion of the inter-frame feature is inherently necessary, we first concatenate the features from the dual streams to get $\mathcal{F}$ and then adopt the refiner [32] to capture the relationship between local features, which helps to perceive the attribution of the local details of the deformed garments, even on the extreme samples. Finally, this module outputs the PC NOCS P_{k+1}^*, which can be used for $(k+1)$-th frame garment pose represented in NOCS.

4.2. Voting-based Projection

In $(k+1)$-th frame, to conduct deformation in 2D plane with the given refined PC NOCS P_{k+1}^* and the mech NOCS P_{k+1}', our key idea can be summarized as follows: reformulating this problem as a projection task and learning semantically 2D warping vector. The key idea behind our

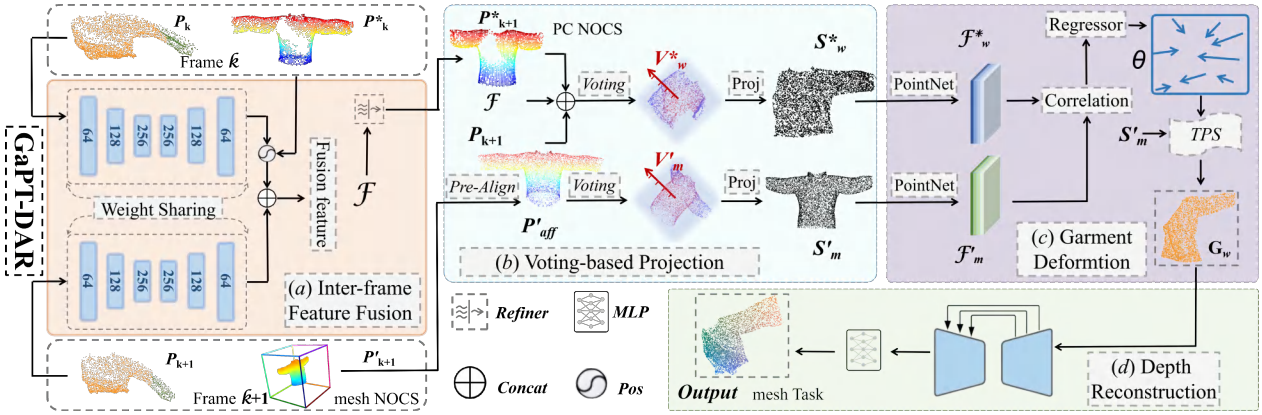


Figure 2. **The overview of our GaPT-DAR**. Formally, taking *partial* observation P_k in k-th frame and P_{k+1} in $(k+1)$-th frame as input, using PC NOCS P_k^* in k-th frame and mesh NOCS P'_{k+1} in $(k+1)$-th frame as auxiliary information, we output the *complete* $(k+1)$-th frame observation (mesh Task) in task space. GaPT-DAR consists of the following components: **(a) Inter-frame Feature Fusion** from input used for the downstream network (Section 4.1). **(b) Voting-based Projection.** A voting-based mechanism is proposed to conduct 3D-2D projection via the optimal projection plane (Section 4.2). **(c) Garment Deformation.** We perform garment deformation guided by TPS transformation parameter θ. (Section 4.3). **(d) Depth Reconstruction.** We recover the depth for point sets G_W and output the complete point cloud (mesh Task) in the $(k+1)$-th frame (Section 4.4).

method is to determine a projection plane to conduct projection from 3D coordinates to a 2D plane. However, determining an optimal projection plane directly is non-trivial. we notice that its normal vector can really help: the normal vector $\mathbf{u}$ of the optimal plane can be regarded as the vertical axis, and the intersection of the vertical axis and the plane can be considered as the pivot point $\mathbf{q}$. In other words, we define the plane by its axis parameters $\phi = (\mathbf{u}, \mathbf{q})$, where $\mathbf{u} \in \mathbb{R}^3, \mathbf{q} \in \mathbb{R}^3$. The normal vector $\mathbf{u}$ ensures the orientation of the optimal plane, while the pivot point helps to determine the location of the plane.

Since the projection scheme of the dual stream is similar, in what follows, we take the mesh NOCS branch as an example. The fundamental mechanism behind our method is an offset-voting scheme with a heatmap, which aims to predict the pivot point $\mathbf{q}$ and the normal vector $\mathbf{u}$ from $P'_{k+1} = \{p_i\}_{i=1}^N$. Concretely, for each point p_i, we explicitly regress an axis vector $V_i \in \mathbb{R}^7$, The first three dimensions of V_i demonstrate the normal vector of p_i for the optimal projection plane. The second three dimensions represent the offset of p_i to the pivot, and the rest dimension represents the heatmap (This is considered as the probability of a candidate point becoming a pivot point).

For normal vector prediction, we perform a dense regression by building a three-layer MLP and output 3N channels to regress the $\mathbf{u}$ for each point p_i. The final predictions of the $\mathbf{u}$ will be the average prediction over all the points with heat scores larger than the threshold of 0.5. Meantime, for pivot point prediction, we additionally output 4N channels, where N channels indicate the heatmap of p'_i and 3N channels indicate the offset between this point and the GT pivot.

Therefore, given the predicted pivot point $\hat{p} = (\hat{x}, \hat{y}, \hat{z})$ and normal vector $\mathbf{u} = (\alpha, \beta, \gamma)$, we regard it as the vertical axis of the optimal projection plane. The point with the highest probability through the heatmap will be considered as the predicted result of $\mathbf{q}$.

Given $\mathbf{u}$ and $\mathbf{q}$, we conduct the projection procedure as follows: intuitively, due to the free rotation of a plane about its normal vector, projection coordinates become non-unique. Thus, we define a common direction vector $\mathbf{e} = [1, 0, 0]$ for two branches meantime to ensure consistency in their projection interpretation. Mathematically, for each point from $P'_{k+1} = \{p'_i\}_{i=1}^N = \{(x'_i, y'_i, z'_i)\}_{i=1}^N$, the projection procedure can be formulated as:

$$(\mathrm{u}'_i,\ \mathrm{v}'_i) = (\mathbf{d} \cdot \mathbf{e},\ ||\mathbf{d} \times \mathbf{e}||_2) \tag{1}$$

$$\text{where } \mathbf{d} = (p'_i - \hat{p}) \times \frac{\mathbf{u}}{||\mathbf{u}||_2} \tag{2}$$

where $||\mathbf{d}||_2$ is the vertical distance from the point p'_i to the normal vector $\mathbf{u}$. Point sets $\{(\mathrm{u}'_i, \mathrm{v}'_i)\}_{i=1}^N \in S'_m$ form the projection result in the optimal 2D projection plane.

4.3. Garment Deformation

Deformation of the 2D plane has received extensive study in the past literature [11, 15]. Common methods such as affine transformation [42], Homography [7], TPS [33], etc. Among them, Thin Plate Spline (TPS) Transformation is a curve-based algorithm that can adaptively distort different regions [3], while all points must act uniformly or multiple viewpoints are required in other methods. Given these considerations, we propose a 2D TPS-conditioned garment

deformation scheme, which warps the rest state garment $\{(\mathrm{u}_i', \mathrm{v}_i')\}_{i=1}^{N} \in S_m'$ guided by the garment that undergoes deformation $\{(\mathrm{u}_i^*, \mathrm{v}_i^*)\}_{i=1}^{N} \in S_w^*$. Our method aims to explore the distribution of possible task space of the garments in 2D plane, which not only inherits the benefits of the aforementioned deformation methods but provides a simple and straightforward idea of warping.

Since S_m' and S_w^* are a family of two-dimensional coordinate groups, to bridge the domain gap between 3D space and 2D point sets, we use the 2D Pointnet to exact warped features $\mathcal{F}_w^*$ and rest state feature $\mathcal{F}_m'$ from S_w^* and S_m', which is essential and fundamental to the learning of warping vector $\theta \in \mathbb{R}^{N\times 2}$. Later on, $\mathcal{F}_w^*$ and $\mathcal{F}_m^{'}$ are adopted as inputs and fed into the correlation layer to calculate the matching score (element-wise multiplication). After that, a regressor is used to predict the warping vector θ (TPS transformation parameter) to model the garment deformation explicitly. However, since there is a giant gap between the garment in the rest state (S_m') and the deformed garment (S_w^*), we find that estimating θ directly is non-trivial. To this end, we perform a self-adaptive pre-alignment procedure before the projection for P_{k+1}', which aims to transform it to the proper position and scale before our TPS transformation. This procedure can be formulated as an affine transformation:

$$P_{aff}' = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{bmatrix} (P_{k+1}')^{\mathrm{T}} - \begin{bmatrix} s\cdot x' - x^* \\ s\cdot y' - y^* \\ s\cdot z' - z^* \end{bmatrix} \quad (3)$$

where P_{aff}' denotes the transformed garment item (show in Fig. 2). (x', y', z') and (x^*, y^*, z^*) are used to represent the center of P_{k+1}' and P_{k+1}^* respectively. We use s as a rescaling factor computed by comparing the aspect ratio to ensure that the aligned garment is almost equal to the warped garment.

Here, we share a more generalized expression to understand the above process: we start by aligning the centers of P_{k+1}' and P_k^* and roughly re-size them to a uniform size, which can facilitate the TPS warping process and help the warping vector θ learn a well-modulated feature.

When it comes to the TPS transformation, we apply the feature $\mathcal{F}_m^{'}$ from the rest state and the pose-aware representation $\mathcal{F}_w^*$ into the geometric matching network, which conducts the regression of TPS parameters θ. Given the $\theta \in \mathbb{R}^{N\times 2}$, we conduct warping process by mapping $S_m^{'} \in \mathbb{R}^{N\times 2}$ to $S_w^* \in \mathbb{R}^{N\times 2}$. By this way, a better representation of θ facilitates us to warp S_m' to the warped garment S_w^*. The errors between S_m' and GT S_w^* can be defined as Garment Deformation loss $\mathcal{L}_{\mathrm{GD}}$:

$$\mathcal{L}_{\mathrm{GD}} = \sum_i \|\nabla(T(S_m'|\theta) - S_w^*)\|_1 \quad (4)$$

where $T(\cdot|\theta)$ denotes the TPS transformation with θ.

4.4. Depth Reconstruction

Interpolation-based methods are commonly employed in 3D surface reconstruction tasks. While these methods can help mitigate missing information by mapping partial point clouds to complete ones, they often struggle to efficiently recover detailed information such as wrinkles and bumps. Based on this consideration, how might the network focus on the depth vector of each point, disentangle it from the latent feature space, and subsequently recover the depth information? This perspective derives our strategy. Our solution concerns the following requirements: (1) **Consistency.** The reconstruction quality is initially assessed by focusing on surface consistency, ensuring smooth and coherent surfaces. (2) **Gradient.** In practice, we find it's not enough to learn a good depth map considering these requirements. [29] prompts us that the normal maps outperform depth maps in perceiving more detailed geometric information. The specific details of our scheme are introduced as follows:

The predicted warped garments G_w will be fed in a convolutional layer serving as a regressor, which aims to output the depth map. We use 3 customized losses to meet the aforementioned requirements: Firstly, to guarantee consistency, we change the vanilla L1 loss to be Log-L1 version motivated by [13]. This helps our network to penalize low-frequency differences between the estimated and the ground truth depth map, resulting in a relatively smooth depth result. The consistency loss $\mathcal{L}_{\mathrm{c}}$ is formulated as:

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{N}\sum_{i=1}^{N} \ln\left(||\hat{d}_i - \widetilde{d}_i||_1 + 1\right) \quad (5)$$

where $\hat{d}_i$ and $\widetilde{d}_i$ are the predicted depth and GT depth of the i-th point respectively, and N is the total number of the valid depth map points.

Afterward, to get a trade-off between the complexity and effectiveness of our framework, we focus on network optimization rather than blindly stacking network modules and deeper networks. We conjecture that depth gradient can really help. To recover the subtle geometric details and further strengthen the depth estimation, a depth gradient loss $\mathcal{L}_{\mathrm{g}}$ is harnessed to enhance the learning of depth-aware features which is defined as:

$$\mathcal{L}_{\mathrm{g}} = \frac{1}{N}\sum_{i=1}^{N} \left(\ln\left(\diamondsuit_x\left(\xi_i\right)+1\right) + \ln\left(\diamondsuit_y\left(\xi_i\right)+1\right)\right) \quad (6)$$

where $\diamondsuit$ denotes the Sobel operator, ξ_i denotes the absolute error of the depth prediction of i-th point. We can obtain normal maps from depth gradient maps referred to [22], whose difference can also be penalized by Eq.6. In practice, we observe that the depth gradient from gullies and

hills-liked regions have significantly larger gradient values, which are constrained along the normal direction and help recover geometric details.

The above loss functions reinforce and complement each other, and are used to constrain different errors in the depth reconstruction process: (1) $\mathcal{L}_{\mathrm{c}}$ is fully exploited to enhance the consistent information along the optimal vertical axis aforementioned in Sec. 4.2. (2) With the auxiliary of the normal direction, $\mathcal{L}_{\mathrm{g}}$ drags the attention on the optimal projection plane. Finally, we utilize a convex combination of these loss functions and the final loss for depth reconstruction module $\mathcal{L}_{\mathrm{DR}}$ is:

$$\mathcal{L}_{\mathrm{DR}} = \lambda_c \mathcal{L}_{\mathrm{c}} + \lambda_g \mathcal{L}_{\mathrm{g}} \tag{7}$$

where λ_c and λ_g are weight factors.

5. Experiments

5.1. Experimental Setup

Dataset. We have evaluated our method on **VR-Folding** dataset made by [32] for category-level garments pose estimation task. This dataset contains two garment manipulations *flattening* and *folding* that consider garments under the off-body environment. There are four garment categories, *i.e*. Shirt, Pants, Top, and Skirt, in VR-folding, which are derived from CLOTH3D[1].

Implementation and Metrics. In this work, experiments are implemented by PyTorch and Two NVIDIA TESLA T4 GPUs. when dealing with the input partial point cloud and the input canonical mesh, we randomly sample 4000 points from the former and 6000 points from the latter's surface for each frame. The number of total training epochs is 200. **NOCS Coordinate Distance** (D_{nocs})), **Chamfer Distance** (D_{champ})), and Correspondence Distance (D_{corr}), A_d). are used as metrics.

Baselines. Treat [8] and [32] as the baselines, We conducted comparative experiments in the following methods or settings: (1) **GarmentNets** [8] is the prior art for category-level garment pose estimation, We adapt it into the tracking task by frame-wise prediction. (2) **Garment-Tracking.** Going beyond GarmentNets and in a similar spirit, [32] introduced the garment into video streaming for observation and performed the first category-level garment Tracking task.

5.2. Comparison with the SOTA Methods

5.2.1. Quantitative Results

For quantitative evaluation, we conduct experiments on the VR-Folding dataset (*Folding* and *Flattening* task). Table 1 reports the performance in detail. See them all, our method achieves better performance on all metrics compared to other baseline methods. Treat them equally: the most difficult metric is A_{3cm} in *Folding* task and A_{5cm} in *Flattening* task, which is hard to cope with due to the narrow error tolerance. Our method outperforms GarmentNets in this metric by a large margin and is also capable of beating GarmentTracking with an average of **2%** superior performance (e.g., **29.8%** vs **32.3%** in Shirt Folding). Moreover, we also do better in both pose estimation and surface reconstruction tasks compared to other methods, which can be proved by mean correspondence distance D_{corr} and Chamfer distance D_{champ}. we conjecture this is attributed to our use of the depth reconstruction module and 3 customized losses, which can perceive the point-wise geometric details.

5.2.2. Qualitative Results

Qualitative results are shown in Fig. 3. Comparing *Folding* task and *Flattening* task, all the methods prefer the latter and show better performance under the same settings. Severe self-occlusion should be responsible for this, which leads to the giant difficulty in reconstructing the complete cloud. However, our method can also excel in this case compared with all the current methods, which can perceive the full configuration from partial visual observation better. Besides, we find that the predictions from GarmentNets are barely adequate for garment deformation perception in continuous frames since its unique framework can't fuse information between frames, instead only reconstructing the video stream frame by frame individually. GarmentTracking, while capable of outputting relatively accurate results, suffers from some significant errors due to its implicit warping methods. Our results are closest to GT over the full range of frames, which echoes the quantitative results (e.g. $D_{nocs} = 0.098$ for ours *vs*. $D_{nocs} = 0.109$ for GarmentTracking for *Shirt Flattening* in Table 1).

5.3. Ablation Analysis

Projection Mechanism. A better projection plane leads to less loss of information (In this paper, it refers to the smaller overlap created by the projection process). In this section, we consider two other projection mechanisms: (1) linear projection using SVD algorithm. (2) xOz plane that is the best projection plane with our observation. Table 2 shows the comparison results. Our method maintains the best performance in all metrics. We conjecture the conclusion as follows: SVD is just a purely linear probing algorithm that does not efficiently capture the semantically optimal projection plane, thus losing a lot of information and bringing in additional noise, while xOz plane can only consider projection relations in a single direction, which leads to undesirable results since it is non-parametric.

3D *vs*. 2D Prediction for Deformation Function. As mentioned before, the fundamental mechanism of our

Type	Method	Folding					Flattening				
		$A_{3cm}\uparrow$	$A_{5cm}\uparrow$	$D_{corr}\downarrow$	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$	$A_{5cm}\uparrow$	$A_{10cm}\uparrow$	$D_{corr}\downarrow$	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$
Shirt	GarmentNets [8]	0.8%	21.5%	6.40	1.58	0.221	13.2%	59.4%	10.54	3.54	0.135
	GarmentTracking[32]	29.8%	85.8%	3.88	1.16	0.051	30.7%	83.4%	8.63	1.75	0.105
	GaPT-DAR (Ours)	**32.3%**	**89.3%**	**3.85**	**1.03**	**0.044**	**33.1%**	**87.6%**	**8.61**	**1.68**	**0.098**
Pants	GarmentNets [8]	16.2%	69.5%	4.43	1.30	0.162	1.5%	42.4%	12.54	4.19	0.185
	GarmentTracking[32]	47.3%	94.0%	3.26	1.07	0.039	31.3%	78.2%	8.97	1.64	0.113
	GaPT-DAR (Ours)	**49.1%**	**95.2%**	**3.15**	**0.99**	**0.031**	**32.8%**	**79.3%**	**8.79**	**1.44**	**0.101**
Top	GarmentNets [8]	10.3%	53.8%	5.19	1.51	0.148	21.6%	57.6%	9.98	2.13	0.174
	GarmentTracking[32]	37.9%	85.9%	3.75	0.99	0.051	36.5%	69.0%	9.41	1.59	0.113
	GaPT-DAR (Ours)	**39.5%**	**88.3%**	**3.59**	**0.89**	**0.045**	**37.6%**	**73.2%**	**9.38**	**1.48**	**0.107**
Skirt	GarmentNets [8]	1.1%	30.3%	6.95	1.89	0.239	0.1%	7.9%	18.48	5.99	0.287
	GarmentTracking[32]	23.8%	71.3%	4.61	1.33	0.060	5.4%	39.4%	16.09	2.02	0.199
	GaPT-DAR (Ours)	**25.1%**	**73.1%**	**4.55**	**1.25**	**0.048**	**8.0%**	**39.9%**	**15.98**	**1.99**	**0.187**

Table 1. **Quantitative results for category-level garments pose tracking on VR-Folding dataset.** There are four garment categories (Shirt, Pants, Top and Skirt) and two garment manipulation tasks (Folding and Flattening).

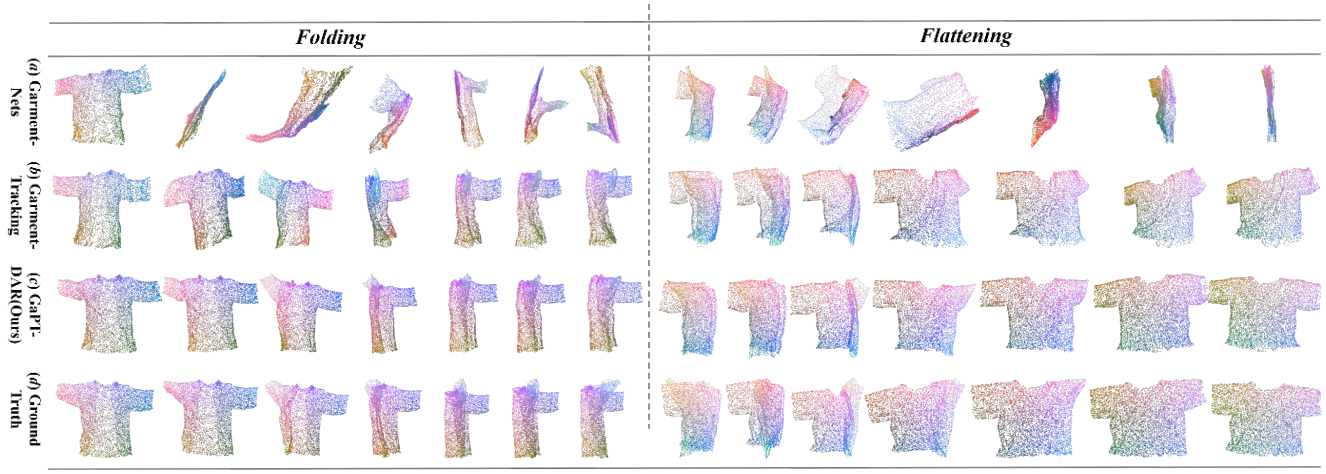


Figure 3. **The qualitative results on category-level garment pose tracking task from the VR-Folding dataset.** We illustrate pose tracking performance on folding (left) and flattening (right) garment manipulations. In the long sequence tracking, our prediction could recover more geometric details which still keeps more in step with GT compared to state-of-the-arts.

Projection Method	Folding		Flattening	
	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$
SVD-based	1.53	0.119	2.50	0.247
xOz plane	1.73	0.136	2.78	0.287
Voting-based	**0.99**	**0.031**	**1.44**	**0.101**

Table 2. **Comparison of different projection methods.** Here we only report experimental results on Pants due to the limited space.

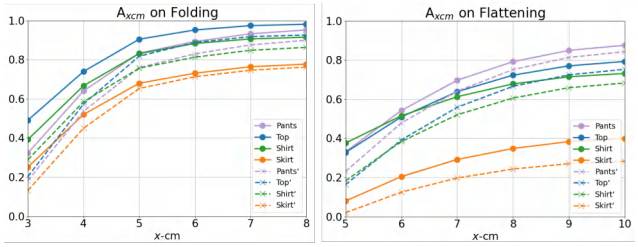


Figure 4. **Comparison between direct 3D and integrated 3D-2D deformation learning.** The solid lines denote the 3D-2D manner while the dotted lines denote the direct 3D manner.

method is the 3D-2D explicit warping procedure and 2D-3D point-wise depth reconstruction. In this part, we conduct 3D prediction for deformation function directly. Concretely, after feature fusion, we use a 3D GNN [24] to regress the coordinates of the complete point cloud in NOCS directly. Comparison results are shown in Fig. 4. It shows that direct 3D manner can't perform well, especially showing a larger performance gap in *Flattening* task. This can be concluded that direct regression of 3D coordinates might not perceive the details of the warping process well.

Noise. The first-frame garment pose and the input canonical shape are prone to be perturbed with noise when faced with

Noise Level	$\mathbf{s}_{pc}$	$\mathbf{o}_{pc}$	δ	$\mathbf{s}_{mesh}$
1x	$[0.8, 1.2]^3$	$[0, 0.1]^3$	0.05	$[0.8, 1.2]^3$
2x	$[0.6, 1.4]^3$	$[0, 0.2]^3$	0.10	$[0.6, 1.4]^3$
3x	$[0.4, 1.6]^3$	$[0, 0.3]^3$	0.15	$[0.4, 1.6]^3$

Table 3. **The noise parameters in the robustness experiment**. $[a, b]^3$ indicates a 3-D vector (*i.e.* x, y, z axis) in which each dimension is uniformly sampled from $[a, b]$.

Pertubation		GarmentTracking [32]		GaPT-DAR (Ours)			
		Flattening	Folding	Flattening		Folding	
★	0×	0.105	0.039	**0.098**		**0.031**	
	1×	0.113	0.039	**0.109**	**0.004**↓	**0.035**	**0.004**↓
	2×	0.158	0.083	**0.151**	**0.007**↓	**0.078**	**0.005**↓
	3×	0.172	0.093	**0.161**	**0.011**↓	**0.087**	**0.006**↓
♣	1	0.105	0.039	**0.098**		**0.031**	
	1/2	0.110	0.052	**0.106**	**0.004**↓	**0.049**	**0.003**↓
	1/4	0.151	0.064	**0.145**	**0.006**↓	**0.058**	**0.006**↓
	1/6	0.173	0.076	**0.164**	**0.009**↓	**0.071**	**0.005**↓
	1/8	0.187	0.088	**0.178**	**0.009**↓	**0.078**	**0.010**↓

Table 4. **Robustness evaluation results.** We test the effects of noisy pose initialization and frame interval for our GaPT-DAR. Note that we use ★ to represent Noise Level, while ♣ is used to represent Frame Keep Ratio. The metric is D_{nocs}.

practical application, so we conduct robustness experiments regarding additional noise distribution following [32]. Concretely, a global scaling factor, a global offset, and a Gaussian noise standard deviation are performed on the coordinates of PC NOCS and mesh NOCS from the first frame. The detailed setting of these noise parameters is shown in Table 3. Quantitative results are shown in Table 4 (Top), our method is more robust than GarmentTraking under different noise levels. Note that when posting different methods to garment perception for embodied AI, noise can frequently occur. Our model performs better on heavy noise scenarios especially, we infer that this is attributed to our explicit warping method and point-aware depth reconstruction.

5.4. Robustness Analysis

Large Frame Interval. In this part, we conduct robustness evaluation under a large frame interval. Concretely, we downsample the number of frames in the video stream so that the number of frames decreases to 1/2, 1/4, 1/6, and 1/8 of the original while keeping the duration constant, and we promise that this downsampling process is random. The robustness study results are displayed in Table 4 (Down). Under all frame rates, our method demonstrates greater robustness. Additionally, our method shows a more robust performance on *Folding* task against missing frames compared with *Flattening* task.

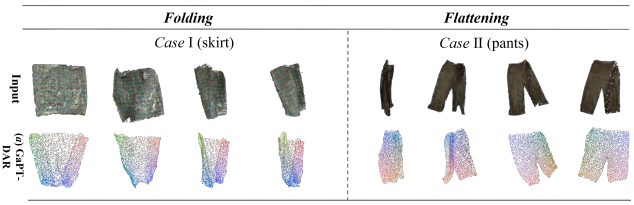


Figure 5. **The qualitative results on unseen instances for *Folding* and *Flattening* task in real-world data.**

5.5. Sim-to-Real Transfer

To assess the generalization capability of the GaPT-DAR, we conduct training on synthetic models and subsequently evaluate its performance on real-world scenarios. For this purpose, we employ a real-world dataset sourced from [32]. The results, presented in Fig. 5, demonstrate the effectiveness of our approach in accurately tracking garment poses for novel garments under real-world conditions.

6. Limitations

Firstly, under severe self-occlusion conditions, where occluded parts are imperceptible (i.e., their pose cannot be captured), our method results in degraded performance. Secondly, despite promising pose tracking results with real-world data, our method necessitates a substantial amount of meticulously labeled high-quality data and meticulously aligned 3D meshes, which are resource-intensive and expensive. Future work should explore self- or weakly-supervised methods to alleviate this reliance on costly data and address this limitation effectively.

7. Conclusion

In this paper, the proposed GaPT-DAR framework that solves category-level garment pose tracking task via integrated 2D deformation and 3D reconstruction. Our method designs a Voting-based Projection, Garment Deformation module, and Depth Reconstruction module to achieve the 3D-2D-3D pose learning pipeline. Experimental results show that our GaPT-DAR is both quantitatively and qualitatively better than state-of-the-arts. Future works will consider practical applications that address the simulation and real-world gap issue, as well as pose tracking under egocentric manipulation of embodied agents.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62302143, in part by Anhui Provincial Natural Science Foundation under Grant 2308085QF207, and in part by National Key Research and Development Program of China under Grant 2021YFD2000201.

References

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

[2] 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] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989.

[4] 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.

[5] Robert Bridson, Sebastian Marino, and Ronald Fedkiw. Simulation of clothing with folds and wrinkles. In *ACM SIGGRAPH 2005 Courses*, pages 3–es. 2005.

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *arXiv preprint arXiv:2309.17448*, 2023.

[7] Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9833–9842, 2023.

[8] Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3324–3333, 2021.

[9] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8958–8966, 2019.

[10] 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.

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

[12] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. *Advances in Neural Information Processing Systems*, 34:27940–27951, 2021.

[13] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019.

[14] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 18–35. Springer, 2020.

[15] Gabriele La Valle and Sina Massoumi. A new deformation measure for micropolar plates subjected to in-plane loads. *Continuum Mechanics and Thermodynamics*, pages 1–15, 2022.

[16] Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. Chord: Category-level hand-held object reconstruction via shape deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9444–9454, 2023.

[17] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020.

[18] 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.

[19] Liu Liu, Jianming Du, Hao Wu, Xun Yang, Zhenguang Liu, Richang Hong, and Meng Wang. Category-level articulated object 9d pose estimation via reinforcement learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 728–736, 2023.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[21] 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.

[22] Yosuke Nakagawa, Hideaki Uchiyama, Hajime Nagahara, and Rin-Ichiro Taniguchi. Estimating surface normals with depth image gradients for fast and accurate registration. In *2015 International Conference on 3D Vision*, pages 640–647. IEEE, 2015.

[23] 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.

[24] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1711–1719, 2020.

[25] 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.

[26] Thomas Stumpp, Jonas Spillmann, Markus Becker, and

Matthias Teschner. A geometric deformation model for stable cloth simulation. *VRIPHYS*, 8:39–46, 2008.

[27] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020.

[28] 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.

[29] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *European Conference on Computer Vision*, pages 430–446. Springer, 2020.

[30] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13209–13218, 2021.

[31] Haoyu Xiong, Haoyuan Fu, Jieyi Zhang, Chen Bao, Qiang Zhang, Yongxi Huang, Wenqiang Xu, Animesh Garg, and Cewu Lu. Robotube: Learning household manipulation from human videos with simulated twin environments. In *Conference on Robot Learning*, pages 1–10. PMLR, 2023.

[32] 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.

[33] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating preserving image content. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7847–7856, 2020.

[34] 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.

[35] Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6875, 2022.

[36] Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. *arXiv preprint arXiv:2309.16264*, 2023.

[37] 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*, pages 5504–5514, 2019.

[38] Fan Zhang and Yiannis Demiris. Learning garment manipulation policies toward robot-assisted dressing. *Science robotics*, 7(65):eabm6010, 2022.

[39] Li Zhang, Zean Han, Yan Zhong, Qiaojun Yu, Xingyu Wu, et al. Vocapter: Voting-based pose tracking for category-level articulated object via inter-frame priors. In *ACM Multimedia 2024*, 2024.

[40] Li Zhang, Yan Zhong, Jianan Wang, Zhe Min, Liu Liu, et al. Rethinking 3d convolution in ℓ_p-norm space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[41] Li Zhang, Weiqing Meng, Yan Zhong, Bin Kong, Mingliang Xu, Jianming Du, Xue Wang, Rujing Wang, and Liu Liu. U-cope: Taking a further step to universal 9d category-level object pose estimation. In *European Conference on Computer Vision*, pages 254–270. Springer, 2025.

[42] Zhimeng Zhang and Yu Ding. Adaptive affine transformation: A simple and effective operation for spatial misaligned image generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1167–1176, 2022.

[43] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13239–13249, 2021.