UniGarmentManip: A Unified Framework for Category-Level Garment Manipulation via Dense Visual Correspondence

Ruihai Wu[*1,4]   Haoran Lu[*2,1]   Yiyan Wang[3,1]   Yubo Wang[2,1]   Hao Dong[†1,4]

[1]CFCS, School of CS, PKU   [2]School of EECS, PKU   [3]School of CS&T, BIT

[4]National Key Laboratory for Multimedia Information Processing, School of CS, PKU

Abstract

Garment manipulation (e.g., unfolding, folding and hanging clothes) is essential for future robots to accomplish home-assistant tasks, while highly challenging due to the diversity of garment configurations, geometries and deformations. Although able to manipulate similar shaped garments in a certain task, previous works mostly have to design different policies for different tasks, could not generalize to garments with diverse geometries, and often rely heavily on human-annotated data. In this paper, we leverage the property that, garments in a certain category have similar structures, and then learn the topological dense (point-level) visual correspondence among garments in the category level with different deformations in the self-supervised manner. The topological correspondence can be easily adapted to the functional correspondence to guide the manipulation policies for various downstream tasks, within only one or few-shot demonstrations. Experiments over garments in 3 different categories on 3 representative tasks in diverse scenarios, using one or two arms, taking one or more steps, inputting flat or messy garments, demonstrate the effectiveness of our proposed method. Project page: https://warshallrho.github.io/unigarmentmanip.

1. Introduction

Next-generation robots should have the abilities to manipulate a large variety of objects in our daily life, including rigid objects, articulated objects [8] and deformable objects [45]. Compared with rigid and articulated objects, deformable objects are much more difficult to manipulate, for the highly large and nearly infinite state and action spaces, and complex kinematic and dynamics. Garments, such as shirts and trousers, are essential types of deformable objects, due to the potentially wide-range applications for both industrial and domestic scenarios. Manipulating garments, such as unfolding, folding and dressing up, has garnered

significant interest in robotics and computer vision fields.

There have been a long range of studies on manipulating relatively simple shaped deformable objects, such as square-shaped cloths [23, 45, 46], ropes and cables [33, 45, 46], and bags [2, 5]. Nevertheless, manipulating garments presents a substantial challenge, as it necessities the comprehensive understanding of more diverse geometries (garments in a certain category have different shapes, let alone in different categories), more complex states (various geometries with diverse self-deformations), and more difficult goals (*e.g.*, garments require multiple fine-grained actions fold step by step). Many existing studies on garment manipulation rely on large-scale annotated data [1, 4], which is labor-intensive and time-consuming, hindering the scalability in the scenarios of real-world applications. Besides, many works design quite different methods to tackle different specific tasks [1, 4, 41, 50], making it difficult to efficiently share and reuse information among different tasks.

Different from other object types, garments possess a property that, in a certain category, while different garments may have different geometries, they usually share the same structure. For example, tops (such as T-shirts, jackets and jumpers), are composed of certain components (a body with two sleeves and a collar), and the topological structures of the components are usually the same, even though the length, width and geometries of a certain component in different garments may be quite different. Thanks to such similarity in structure shared among garments in the category level, it is easy for humans to fulfill a task on unseen novel garments using the experience of manipulating only one or a few garments in the same category. Therefore, we empower robots with the above one/few-shot generalization ability humans have in diverse tasks, by leveraging such structural similarity among garments.

Among multiple ways to describe and represent garments (*e.g.*, poses [7, 51], lines [10, 58] and keypoints [57]), **skeleton** [34], *i.e.*, a graph of keypoints covering significant points on garment edges and joints to represent the topology of 3D objects, is suitable for describing the above-mentioned structures shared among garments. The skele-
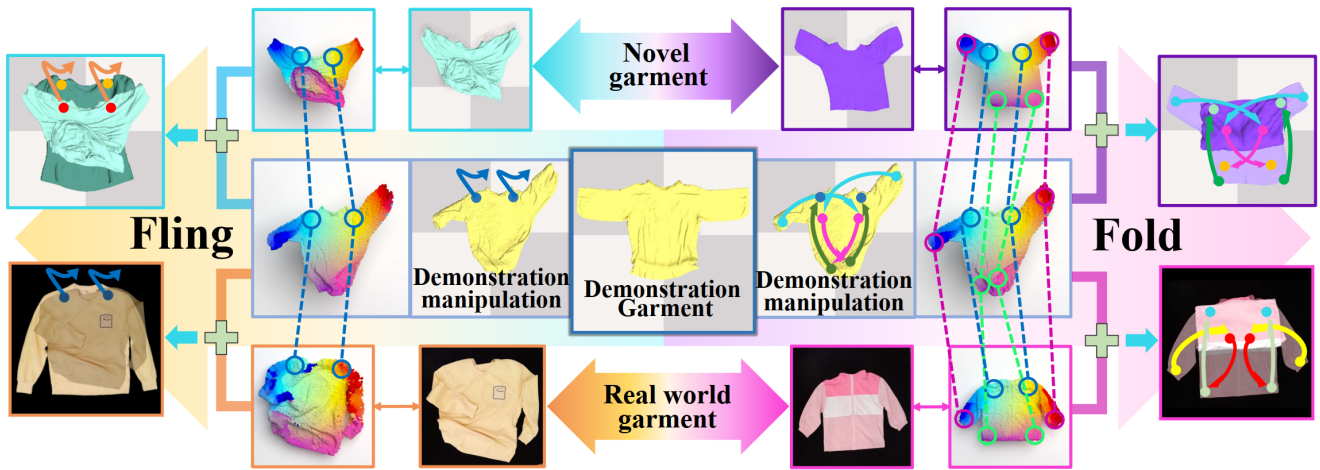

Figure 1. Given a demonstration garment (**Middle**) and the demonstration actions to fulfill a task (**Middle-Left/-Right**), for a novel object, we find the manipulation points using the proposed **Dense Visual Correspondence for Garment Manipulation** and execute the corresponding action to fulfill the task (**Left/Right**). Color similarity denotes in the correspondence space.

ton points are sparse, distinct and ordered, and thus (1) exist on each garment and (2) can easily distinguish with other skeleton points, making them easy to learn. Therefore, we use skeleton points to build structural correspondence among garments. Moreover, as different-extent self-deformations make the garments to be quite complex, while previous works only studied skeleton points on rigid [34], articulated [47] or fixed-posed deformable objects in the canonical view [57], we further extend skeleton points to garments at any deformation states, making a step to more realistic scenarios for garment manipulation.

While skeleton points build topological correspondence between different garments in the skeleton keypoint level, the state and action spaces of garments are exceptionally large and each point on the garment could be the manipulation point, making the sparse skeleton points unable to fully represent garments for manipulation. To represent objects with large state and action spaces, dense (*i.e.*, point-level or pixel-level) object representations, including dense object descriptors [9] and dense visual actionable affordance [28], which indicate the actionable information on each point of the object, have demonstrated its superiority on rigid [9], articulated [43], and simple-shaped deformable object manipulation [45]. We further extend dense object representations to garments, with the awareness of garment correspondences, using the proposed skeleton points, and thus achieve fine-grained manipulation for complicated garments.

With dense visual correspondence aware of garment structures, one demonstration can roughly guide manipulating a novel garment by indicating corresponding action points and policies. Furthermore, as manipulation for specific tasks rely on not only garment structures but also task-specific knowledge, we further transform the representation from task-agnostic structural to task-specific functional for more accurate manipulation in various downstream tasks,

using few-shot demonstrations to achieve this adaptation.

To demonstrate the performance of our proposed representations, we conduct experiments on 3 different kinds of garments over 3 representative tasks. The experimental results showcase the superiority of our proposed framework in manipulating diverse novel garments in multiple tasks using the proposed dense visual correspondence and one or few-shot demonstrations.

In summary, our contributions include:

- We propose to learn category-level dense visual correspondence to reflect the topological and functional correspondence across garments in different styles or deformations, which is an unified representation that facilitates manipulating diverse unseen garments in multiple tasks with one or few-shot demonstrations.
- We propose a novel learning framework with novel designs to efficiently learn the proposed dense visual representation for garments.
- Experiments over diverse representative tasks demonstrate the effectiveness of our proposed dense correspondence and the learning framework.

2. Related Work

2.1. Dense Representations for Manipulation

Dense object descriptors [9] that learn point- or pixel-level object representations are proposed by and for robotic manipulation. A series of works extend such descriptors to propose grasp pose [35, 53], manipulate ropes [36], smooth and fold fabrics [11]. Another series of works learn point-level dense affordance for articulated object [22, 25], deformable object [45], language-guided [48] and bimanual [56] manipulation, as well as exploration for interaction [29, 40], facilitating point-level contact point selection for diverse downstream tasks. Our proposed dense correspondence extends dense object representations in manipulating garments.

2.2. Visual Correspondence Learning

Learning visual correspondence aims to reflect the shared information (*e.g.*, geometric, topological and functional information) between different objects, which facilitates generalization in diverse tasks, including functional perception [19], pose estimation [14], grasping [30, 52, 53] and fabric manipulation [11]. For garments, although different garments have quite different geometries and deformations in different states, they share similar structural and topological information in the category level, which can help in manipulating novel garments with the demonstration of a garment with the similar structure. So we propose to learn correspondence between garments to facilitate novel garment manipulation in diverse downstream tasks.

2.3. Cloth and Garment Manipulation

Manipulating a square-shaped cloth is relatively well-studied, with previous works leveraging flow and dynamics [23, 42], tactile feedback [37, 38], dense representations [33, 45] and reinforcement learning [16, 27] to tackle different tasks. Garment manipulation is more challenging, for the diversity of garment types and shapes, requiring the method to handle diverse objects and states. While previous works mainly learn the policy for a certain task, such as folding [1, 4, 50], unfolding [20], grasping [6, 54] and dressing-up [41, 55], on similar shaped garments, we focus on learning garment representations that can generalize to diverse objects in a category and facilitate many tasks.

3. Problem Formulation

Given an N-point ($N = 10,000$) 3D partial point cloud observation of a garment $O \in \mathbb{R}^{N \times 3}$, garment manipulation aims to manipulate the garment by a sequence of n actions to complete different tasks. As explained in [11, 32, 33], each action a_i includes grasping at a pick point p_{pick_i}, pulling to a place point p_{place_i} without changing the orientation of the end-effector. Additionally, for dual-arm manipulation, each action a_i includes a pair of pick points $p_{pick_i}^1$, $p_{pick_i}^2$ and corresponding place points $p_{place_i}^1$, $p_{place_i}^2$.

Given two garments O_1 and O_2, dense correspondence evaluates the correspondence (normalized to $[-1, 1]$) in topology or function between each point pair (p_1, p_2), with p_1 from O_1 and p_2 from O_2.

Given a task T, a demonstration includes the observation $\hat{O}$ of a demonstration garment, and its corresponding demonstration action (including single and dual-arm actions) sequence $(\hat{p}_{pick_1}, \hat{p}_{place_1}, ..., \hat{p}_{pick_n}, \hat{p}_{place_n})$ that can fulfill T. Given the observation O of a new garment to fulfill T, manipulation using dense correspondence first finds the points $(p_{pick_1}, p_{place_1}, ..., p_{pick_n}, p_{place_n})$ on O, where p_{pick_i} and p_{place_i} have the best correspondence score to $\hat{p}_{pick_i}$ and $\hat{p}_{place_i}$ among all points on O, and then executes the corresponding actions to fulfill the task on O.

4. Method

4.1. Overview

Our framework first learns topological dense visual correspondence aware of different garment deformations and shapes respectively using self-play and skeleton points (Section 4.2), with further coarse-to-fine refinement (Section 4.3). After the few-shot adaptation for different downstream tasks, the learned correspondence turns from *topological* to *functional* (Section 4.4), and thus could facilitate manipulating unseen novel garments on various tasks using one or few-shot demonstrations (Section 4.5). Section 4.6 describes network architectures and the training strategy.

4.2. Self-supervised *Topological* Dense Visual Correspondence Learning

The diversity of garments in different states mainly comes from two perspectives: self-deformations, and styles of objects in the same category. To empower the Dense Visual Correspondence with the alignment ability for different garments in different states, we decouple the learning process into two parts, respectively learning cross-deformation correspondence and cross-object correspondence.

4.2.1 Cross-Deformation Correspondence

Many tasks, such as unfolding and hanging, require manipulating the garment at any random states (*e.g.*, after a random drop). As demonstrated in [11], while garments have complex states and infinite deformations, the manipulation policies (manipulation points) are usually invariant to deformations. To empower the model with the ability to handle garments in different deformations, we introduce learning correspondence across deformations of the same garment.

Given two partial observations O and O' of the same garment in different deformations generated by self-play, and a visible point p on O, we can easily get its corresponding position point p' in O' using point tracing in simulation. If p' is visible, the representations f_p and $f_{p'} \in \mathbb{R}^{512}$ of p and p' extracted by the backbone network $\mathbf{F}$, should be the same, as the representations are agnostic to self-deformations. We normalize point representations to be unit vectors, and thus the similarity between f_p and $f_{p'}$ can be computed by the dot product of f_p and $f_{p'}$, *i.e.*, $f_p \cdot f_{p'}$. For p on O, we use p' on O' as the positive point, and sample m negative points ($m = 150$): $p'_1, p'_2, ..., p'_m$. We pull close f_p and $f_{p'}$, while push away f_p and other point representations. Following InfoNCE [18], a widely-used loss function in one-positive-multi-negative-pair contrastive representation learning, we identify the positive p' amongst m negative samples:

$$\mathcal{L}_{CD} = -log(\frac{exp(f_p \cdot f_{p'}/\tau)}{\sum_{i=1}^{m} exp(f_p \cdot f_{p'_i}/\tau)}) \quad (1)$$

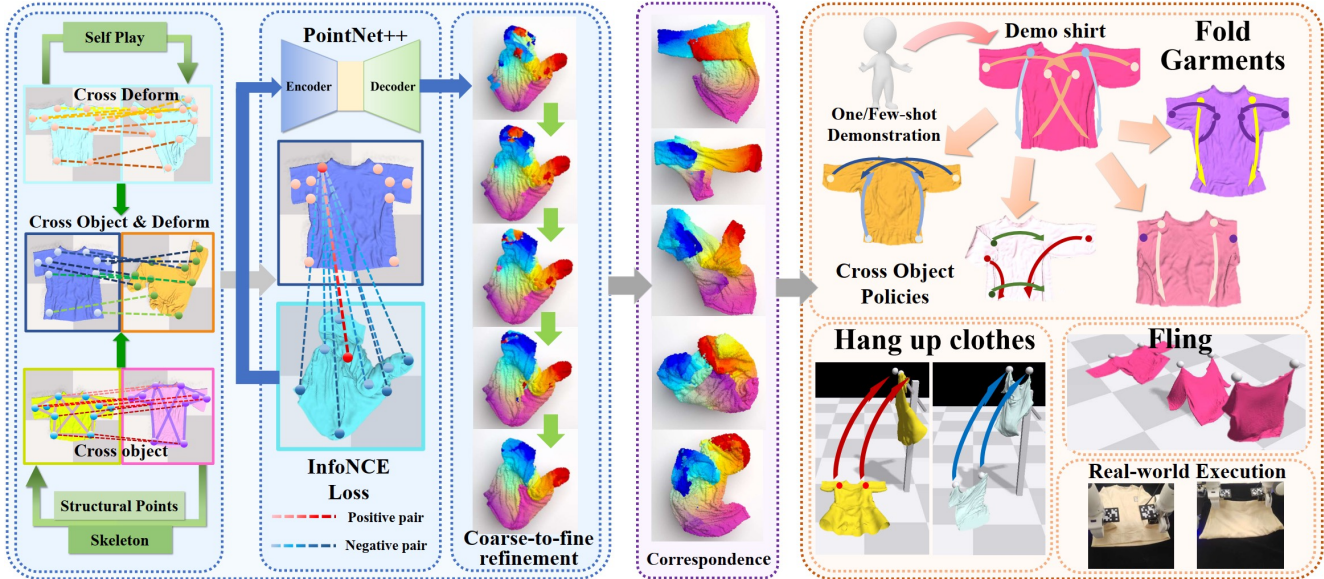, where τ denotes the balancing coefficient in InfoNCE.

Figure 2. **Our Proposed Learning Framework for Dense Visual Correspondence. (Left)** We extract the cross-deform correspondence and cross-object correspondence point pairs respectively using self-play and skeletons, and train the per-point correspondence scores in the contrastive manner, with the Coarse-to-fine module refines the quality. **(Middle)** Learned correspondence demonstrates point-level similarity across different garments in different deformations. **(Right)** The learned point-level correspondence can facilitates multiple diverse downstream tasks using one or few-shot demonstrations.

4.2.2 Cross-Object Correspondence

In a certain category, while garments highly vary in original shapes, such as sizes, length-width ratios, sleeve lengths and styles, they share the same topological structure. The awareness of such structures will make it easy to manipulate unseen novel garments with demonstrations.

To leverage the shared structural information and generalize to novel shapes, we propose to use skeleton, *i.e.*, a graph of keypoints that represents topology of the 3D object, as the shared bridge for different garments with similar structures. The reasons for using skeleton include:

- Skeleton points are **distinct** and **sparse**, thus easy to learn and generalize, compared to complicated representations;
- Skeleton points are **distinct** and **ordered**, making it easy to build topological correspondence between two objects by aligning each specific skeleton point on them;

To learn garment skeletons in the category level, we employ the designs of Skeleton Merger [34], which can generate skeletons for rigid objects, or canonically-posed (*e.g.*, flat) deformable objects. So we generate skeletons for flat garments. Specifically, we generate s ($s \leq 50$) ordered points on the point cloud observation O as skeleton points, with $s \times (s-1)/2$ activation scores $a_{i,j}$ ($1 \leq i,j \leq s$) indicating whether each edge between 2 skeleton points exists on the object. We sample points on each edge, with trainable offset for each sample, merging them into the reconstructed object $\bar{O}$. The training signal is whether $\bar{O}$ cov-

ers O (trains skeleton point positions) and whether sampled points on each edge exists in O (trains $a_{i,j}$), As a result, skeleton points will exist on significant positions on garments (*e.g.*, boundaries, corners and intersections of parts) and meaningful edges will retain, as shown in Figure 2 (the Left-Down part). More implementation details can be seen in the original paper of Skeleton Merger [34].

As skeleton points are **ordered**, given observation O of a flat garment with one of its skeleton point p, we can get the corresponding skeleton point $\tilde{p}$ on the observation $\tilde{O}$ of another flat one, by applying the skeleton network on $\tilde{O}$ and get the skeleton point in the same order of p in O. Then, the topological correspondence between **flat garments** have been built in the **skeleton-point level**. As the features extracted by neural networks are continuous when point positions continuously change, and skeleton points cover the whole garment, the feature of any point can be reflected by its nearby skeleton points (like interpolation) with topological information. Therefore, the representation of each point on the garment will reflect its topology, and **dense correspondence** between flat garments has been naturally built.

4.2.3 Integration of Cross-Deformation and Cross-Object Correspondence

Since we have designed dense correspondence between **the same garment in different deformations**, and dense correspondence between **different flat garments**, the next step

is to aggregate them into one dense representation system on diverse garments in any deformation states.

We first project skeletons of garments in their flat states to any deformation states using point tracing in simulation. Thus, given the observation O in random deformation with one of its skeleton point p, we can get the corresponding skeleton point $\tilde{p}$ on the observation $\tilde{O}$ of another garment in random deformation. If $\tilde{p}$ is visible on $\tilde{O}$, f_p and $f_{\tilde{p}}$ should be the same. For p on O, we use $\tilde{p}$ on $\tilde{O}$ as the positive point, and sample m negative points ($m = 150$): $\tilde{p}'_1$, $\tilde{p}'_2$, ..., $\tilde{p}'_m$. We follow InfoNCE and use $\mathcal{L}_{CO}$ for training:

$$\mathcal{L}_{CO} = -log(\frac{exp(f_p \cdot f_{\tilde{p}'}/\tau)}{\sum_{i=1}^{m} exp(f_p \cdot f_{\tilde{p}'_i}/\tau)}) \quad (2)$$

In the meanwhile, we empower the cross-object correspondence with agnosticism to deformations. We sample the observation O' of the first garment O in another deformation state as described in Section 4.2.1, and train its Cross-Deformation Correspondence using $\mathcal{L}_{CD}$ in Equation 1, together with $\mathcal{L}_{CO}$ in Equation 2 to make the learned representations aware of both cross-deformation and cross-object point-level correspondence.

4.3. Coarse-to-fine Correspondence Refinement

Although above framework can learn the general distributions of all points' representations using offline randomly collected data, some difficult details (such as the boundaries between the folded sleeve on the garment body) should be paid more attention by the model, and there may exist inaccurate representations on some points or areas. The above phenomenon is also demonstrated in previous dense correspondence learning studies for 3D objects [13, 15, 39].

Therefore, we propose the Coarse-to-fine (C2F) Correspondence Refinement procedure to make the model more focused on difficult points on the garment, and eliminate inaccurate predictions, by refining the offline trained model using its online prediction failures.

Specifically, for a certain garment, we sample a point p on the observation O in one deformation state, predict its point-level correspondence score on O' in another deformation state, with p' as corresponding point of p. We collect points $\{p'_1, p'_2, ..., p'_r\}$ that meet the following requirements:
- their correspondence scores are higher than a correspondence thresh α;
- their distances to p' in the canonical (flat) state (denoted as $d_{p'_i}$ for p'_i) are longer than a distance thresh β.

These points are prediction failures of the model trained on offline data. We augment InfoNCE by distance (failure points farther away from p' will receive more penalties to take more focus) to refine the model on them:

$$\mathcal{L}_{C2F} = -log(\frac{exp(f_p \cdot f_{p'}/\tau)}{\sum_{i=1}^{r} d_{p'_i} \cdot exp(f_p \cdot f_{p'_i}/\tau)}) \quad (3)$$

To prevent the model from forgetting the knowledge in offline data, in this procedure, we simultaneously train the model using offline data ($\mathcal{L}_{CO}$ and $\mathcal{L}_{CD}$) and online predictions ($\mathcal{L}_{C2F}$).

4.4. From *Topological* to *Functional*: Few-shot Adaptation for Downstream Tasks

While such topological structure of the above learned correspondence is significantly aligned with cross-object manipulation policy, the point functionalities in different tasks may differ to some extent, and thus could not be adequately reflected by a fixed representation. To adapt the learned topological correspondence to be functional for different downstream tasks, we propose the few-shot adaptation.

With the trained model, for a certain task and a functional action point (such as the pick point on the left sleeve for folding) we annotate l ($l \leq 5$) points $p_1, p_2, ..., p_l$ on l observations $O_1, O_2, ..., O_l$ of different garments in different deformations, and fine-tune the trained model to make $f_{p_1}, f_{p_2}, ..., f_{p_l}$ to be the same using InfoNCE loss similar to Equation 2 (replacing the *topological* correspondence point with the *functional* correspondence point as the positive sample). Consequently, the model can generally keep the topological information while become more aware of the functional information of the certain downstream task.

4.5. Manipulation Policy Generation

As shown in Figure 1 and described in the last paragraph of Section 3, for novel garments over different downstream tasks, we can easily generate manipulation policies by selecting the picking and placing points that are most close to the demonstrations in the correspondence space. More details of policy generation for diverse representative downstream tasks are described in Section 5.4.

4.6. Network Architectures and Training Strategy

Segmentation-version PointNet++ [31] is used as the backbone feature extractor $\mathbf{F}$ that takes the point cloud observation O as input to extract per-point features. The per-point features are directly used to calculate correspondences.

We set batch size to be 32. In each batch, we sample 32 garment pairs. For each garment pair, we sample 20 positive positive point pairs, and 150 negative point pairs for each positive point pair. Therefore, in each batch, $32 \times 32 \times 20$ data will be used to update the model. During the Correspondence training stage, we train the model for 40,000 batches. During Coarse-to-fine Refinement, we train the model for 100 batches. During Few-shot Adaptation, we slightly refine the model using 5 demonstration data.

5. Experiment

5.1. Simulation and Dataset

We build our simulation environment based on the PyFleX bindings [21, 24, 49] to Nvidia FleX [26], equipped with 3 kinds of garments, covering 500 tops (including shirts, hoodies, jumpers and etc.), 600 trousers and 600 dresses with diverse shapes, from the large-scale CLOTH3D dataset [3]. An extra rack is loaded.

5.2. Tasks and Metrics

We evaluate our method over 3 different representative garment manipulation tasks:

- **Unfolding** that unfolds garments at random deformations to be flat. The unfolding succeeds when the coverage area of the unfolded garment exceeds a bar [12]. The two subtasks, **Unfold-RAND** and **Unfold-DROP**, respectively denote the garment initial states are generated by a few random actions or by dropping (more realistic).
- **Folding** that folds garments. A folding succeeds when the Intersection-over-Union (IOU) between the target and the folded garments exceeds a bar [4, 50]. **Fold-FLAT** and **Fold-FLING** respectively denote garment initial states are perfectly flat or generated by flinging (more realistic).
- **Hanging** that hangs garments on the rack, with the successful rate metric [6]. **Hang-RAND** and **Hang-FLING** respectively denote garment initial states are generated by a few random actions or by flinging (more realistic).

5.3. Baselines

For **Folding**, we compare with ClothFunnels [4] that learns keypoints (*e.g.*, endpoints of two sleeves, endpoints of the garment bottom line) from large-scale human-annotated data, pick and place keypoints step by step to fold garments.

For **Hanging**, we compare with GCSR [6] that detects Structural Regions for manipulation (collars for hanging). Besides, we compare with DefoAfford [45] that learns point-level actionable affordance scores for accomplishing the task and selects the best point for interaction.

For **Unfolding**, we compare with FlingBot [12] that predicts the garment coverage area after flinging each two-point grasp pair, and selects the best pair for fling. Besides, we compare with DefoAfford as it demonstrates the capability to unfold fabrics using one gripper.

5.4. Manipulation Policy Generation

For **Unfolding**, we use fling [12] as the action for its quickness in unfolding garments using only a few steps. It picks up 2 points simultaneously with 2 arms to lift the garment, pull the 2 points apart to stretch the garment, fling the garment and place it on the workplace. As some keypoints for flinging may be occluded, we design 4 candidate pick pairs (*e.g.*, the 2 endpoints of the shoulder, the 2 endpoints
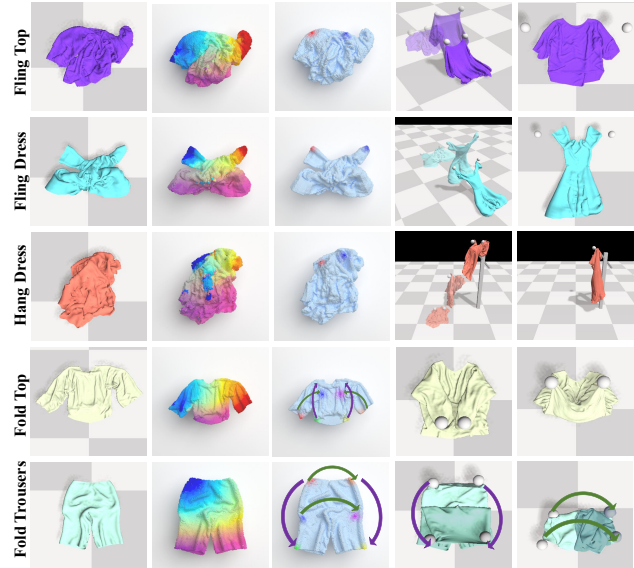

Figure 3. **Correspondence Guided Manipulation on Different Garment Types and Tasks.** From left to right: observation, correspondence, manipulation points (colored points) selected using correspondence to demonstrations and the manipulation action.

of the bottom line), and select pick point pair of the observation with the highest correspondence to designed candidates. We execute the fling action for at most 3 steps.

For **Folding**, we pick and place keypoints (defined in ClothFunnels [4]) on the garments to step by step fold garments with one or two arms. Given the pick and place points sequence in the demonstration, we select their closest pick and place points on each unseen object in the correspondence space and execute the pick-place action sequence. Moreover, to demonstrate the dense representations can facilitate multiple manipulation strategies with slight human annotations, we show 4 folding strategies achieved by our method using one or few-shot demonstrations (Figure 5).

For **Hanging**, we pick the point p_{pick} that is most close to the demonstration pick point $\hat{p}_{pick}$, pull the garment up, and place it on the rack.

5.5. Results and Analysis

Table 1, 2 and 3 present quantitative comparisons with baselines. Figure 4 shows the learned correspondence on different garments shapes and deformations. Figure 3 demonstrates the manipulation actions guided by correspondence.

For folding, as **ClothFunnel**'s keypoint detection model is trained on garments in fully unfolded states, it is difficult to generalize to garments unfolded using FlingBot or our method, as garments could not be perfectly unfolded into a fully flat state. In contrast, as our method is trained with the awareness of self-deformations, it is easier to detect such keypoints in diverse garment states. While **UniFolding**
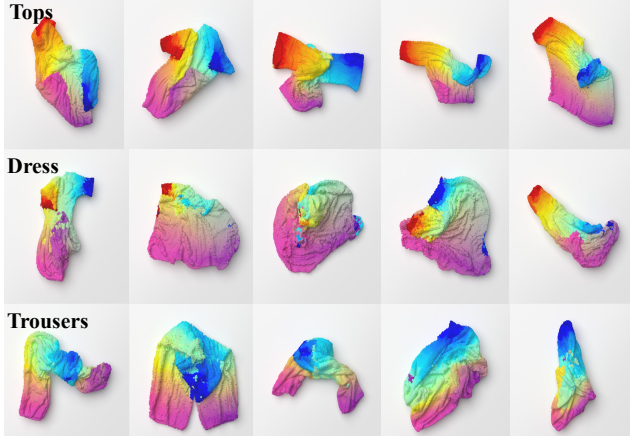
Figure 4. **Learned Dense Visual Correspondence.** For each category, we show correspondence for 5 objects in different deformations. Color similarity denotes correspondence similarity.

Method	Unfold-RAND (%)	Unfold-DROP (%)
FlingBot	80.3 / 82.1 / 79.9	84.3 / 86.7 / 81.9
DefoAfford	70.1 / 63.9 / 61.7	73.8 / 60.9 / 66.7
Ours	**83.6 / 86.9 / 81.6**	**85.3 / 88.1 / 83.6**

Table 1. **Results for Unfolding.** Numbers in the first / middle / last denote results for top / dress / trouser (the same below).

Method	Fold-FLAT (%)	Fold-FLING (%)
ClothFunnels	82.2 / 83.9 / 79.6	61.7 / 63.5 / 60.3
UniFolding	**83.5** / 82.9 / 81.6	**78.7** / 81.5 / 78.6
Ours	**83.5** / 84.0 / 83.3	77.9 / **82.5 / 81.3**

Table 2. **Results for Folding.**

Method	Hang-RAND (%)	Hang-FLING (%)
GCSR	78.5 / 72.3 / 77.9	81.2 / 80.9 / 78.7
DefoAfford	73.4 / 69.2 / 71.6	79.4 / 76.4 / 73.8
Ours	**81.9 / 77.4 / 83.3**	**83.8 / 89.6 / 81.5**

Table 3. **Results for Hanging.**

trains using and thus works well on deformed states, it is designed for the specific folding task. In contrast, our method can facilitate multiple downstream tasks. Furthermore, as shown in Figure 5, our method can work well on different , while policies trained on large-scale annotated data cannot easily generalize to novel manipulation methods.

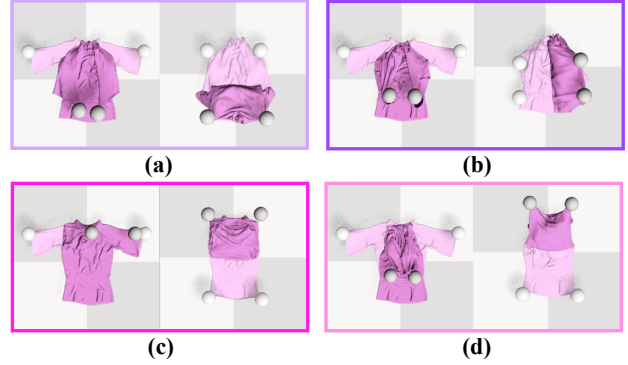For unfolding, DefoAfford cannot perform well in 3

(a)

(b)

(c)

(d)

Figure 5. **Visualization of Different Folding Policies.**

steps as it only utilizes one gripper. For FlingBot, although it is trained using large-scale different states and interactions, without training on many garments, it cannot generalize well to novel garments. Besides, it is costly in time and computing resources, in that it requires separately training 96 models to generate affordance maps in 12 garment rotation types and 8 garment scale types, and then selecting the best pick points pair in all the 96 rotation-scale combinations. In contrast, as the learned dense correspondence is aware of different garment scales and rotations, we can directly use one model to select the grasp points for flinging.

For the comparison with GCSR in folding, as it requires collar detection as Structural Regions, it cannot perform well on garment states where collars are occluded.

It is worth mentioning that, while the baselines are mostly designed for specific tasks and may pose requirements to garment initial states, our proposed dense visual correspondence is an **unified** representation for garments, and thus makes it easy to generate policies for **multiple downstream tasks** on **different garment deformations**.

5.6. Ablation Studies

We compare our method with the ablated versions on the representative folding task, to demonstrate the effectiveness of our method's different components:

- **Ours w/o CO**: our method without learning the cross-object correspondence. This version is similar to [11] that learns correspondence and policies on square-shaped fabrics and similar T-shirts with very short sleeves.
- **Ours w/o CD**: our method without learning the cross-deformation correspondence.
- **Ours w/o C2F**: our method without C2F Refinement.
- **Ours w/o FA**: our method without few-shot adaptation on different downstream tasks.

Table 4 shows quantitative comparisons with ablations. Clearly each component improves our method's capability.

As shown in Figure 6, Coarse-to-fine Refinement eliminates many inaccurate areas of correspondence prediction, and makes the boundaries of garment parts more clear.

Method	Top (%)	Dress (%)	Trouser (%)
Ours w/o CO	75.3	70.8	77.6
Ours w/o CD	63.9	65.2	59.4
Ours w/o C2F	78.0	80.3	76.7
Ours w/o FA	81.9	82.3	78.6
Ours	**83.6**	**86.9**	**81.6**

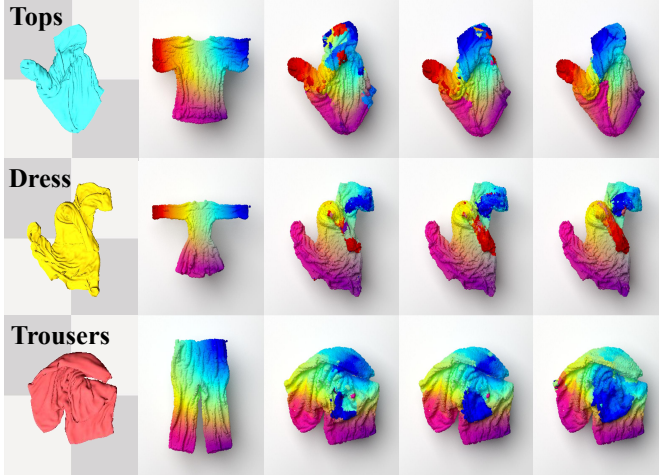Table 4. **Ablation Studies.**

Figure 6. **Coarse-to-fine Refinement**. From left to right: the garment, correspondence in its flat pose, initial correspondence, correspondence after 50 and 100 refinement batches.

Figure 7. **Ablation on Few-shot Adaptation.**

As shown in Figure 7, after few-shot adaptation, the manipulation points tend to be more functional, and thus the folded garments become more organized.

5.7. Real-world Evaluation

Setup. As shown in Figure 8, our real-world experiment setup consists of two Franka Panda robot arms, and a Microsoft Azure Kinect camera (which has demonstrated high-precision with slight noises for robotic manipulation [29, 44]) capturing top-down point cloud. We use Segment Anything (SAM) [17] to segment the garment from the scene and project the segmented image with depth to point cloud.

Please refer to the supplementary materials for more details and videos of real-world manipulations.

To align the scanned point cloud with those in simulation, we annotate a few skeleton points on 3 real-world gar-
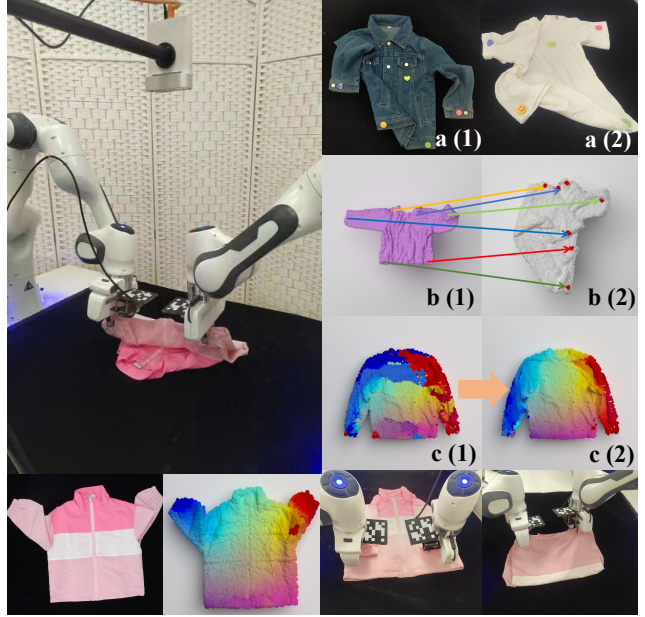

Figure 8. **Real-world Setup and Experiments.**

Method	ClothFunnel	UniFolding	Ours
Fold	8 / 15	10 / 15	**11** / 15

Table 5. **Real-world Evaluation on Folding.**

ments (Figure 8, a), with the correspondence between annotated skeleton points and those in simulation shown in b(1) and b(2), and fine-tune the pre-trained model by pushing close the representations of skeleton points on garments in simulation and the real world using InfoNCE. The correspondence is adapted from c(1) to c(2) in Figure 8.

We use 5 real world different-shaped tops, each conducting folding on 3 different initial deformations, and report the number of successful executions in Table 5.

6. Conclusion

We propose to learn dense visual correspondence for diverse garment manipulation tasks with category-level generalization using only a few annotations. We first train topological correspondence self-supervisedly using self-play and garment skeletons, and then fine-tune it using few-shot demonstrations to transform the topological correspondence to be functional to different downstream tasks. Extensive experiments demonstrate the superiority of our method.

7. Acknowledgment

References

[1] Yahav Avigal, Lars Berscheid, Tamim Asfour, Torsten Kröger, and Ken Goldberg. Speedfolding: Learning efficient bimanual folding of garments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2022. 1, 3

[2] Arpit Bahety, Shreeya Jain, Huy Ha, Nathalie Hager, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Bag all you need: Learning a generalizable bagging strategy for heterogeneous objects. *IROS*, 2023. 1

[3] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 6

[4] Alper Canberk, Cheng Chi, Huy Ha, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation. In *International Conference of Robotics and Automation (ICRA)*, 2022. 1, 3, 6

[5] Lawrence Yunliang Chen, Baiyu Shi, Daniel Seita, Richard Cheng, Thomas Kollar, David Held, and Ken Goldberg. Autobag: Learning to open plastic bags and insert objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3918–3925. IEEE, 2023. 1

[6] Wei Chen, Dongmyoung Lee, Digby Chappell, and Nicolas Rojas. Learning to grasp clothing structural regions for garment manipulation tasks. *arXiv preprint arXiv:2306.14553*, 2023. 3, 6

[7] Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[8] Yushi Du, Ruihai Wu, Yan Shen, and Hao Dong. Learning part motion of articulated objects using spatially continuous neural implicit representations. In *British Machine Vision Conference (BMVC)*, 2023. 1

[9] Peter Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *Conference on Robot Learning*, 2018. 2

[10] Antonio Gabas and Yasuyo Kita. Physical edge detection in clothing items for robotic manipulation. In *2017 18th International Conference on Advanced Robotics (ICAR)*, pages 524–529. IEEE, 2017. 1

[11] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph E Gonzalez, Nawid Jamali, et al. Learning dense visual correspondences in simulation to smooth and fold real fabrics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11515–11522. IEEE, 2021. 2, 3, 7

[12] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning*, pages 24–33. PMLR, 2022. 6

[13] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 5

[14] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6758, 2022. 3

[15] Ling Hu, Qinsong Li, Shengjun Liu, and Xinru Liu. Efficient deformable shape correspondence via multiscale spectral manifold wavelets preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14536–14545, 2021. 5

[16] Rishabh Jangir, Guillem Alenya, and Carme Torras. Dynamic cloth manipulation with deep reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4630–4636. IEEE, 2020. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 8

[18] Cheng-I Lai. Contrastive predictive coding based feature for automatic speaker verification. *arXiv preprint arXiv:1904.01575*, 2019. 3

[19] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021. 3

[20] Yinxiao Li, Danfei Xu, Yonghao Yue, Yan Wang, Shih-Fu Chang, Eitan Grinspun, and Peter K Allen. Regrasping and unfolding of garments using predictive thin shell modeling. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1382–1388. IEEE, 2015. 3

[21] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *International Conference on Learning Representations*, 2019. 6

[22] Yu Li, Xiaojie Zhang, Ruihai Wu, Zilong Zhang, Yiran Geng, Hao Dong, and Zhaofeng He. Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. *arXiv preprint arXiv:2403.02604*, 2024. 2

[23] Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Conference on Robot Learning*, 2021. 1, 3

[24] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, pages 432–448. PMLR, 2021. 6

[25] Suhan Ling, Yian Wang, Ruihai Wu, Shiguang Wu, Yuzheng Zhuang, Tianyi Xu, Yu Li, Chang Liu, and Hao Dong. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. *ICRA*, 2024. 2

[26] Miles Macklin, Matthias Müller, Nuttapong Chentanez, and Tae-Yong Kim. Unified particle physics for real-time applications. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 6

[27] Jan Matas, Stephen James, and Andrew J Davison. Sim-to-real reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, pages 734–743. PMLR, 2018. 3

[28] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 2

[29] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 8

[30] Timothy Patten, Kiru Park, and Markus Vincze. Dgcm-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping. *Frontiers in Robotics and AI*, 7:120, 2020. 3

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[32] Daniel Seita, Aditya Ganapathi, Ryan Hoque, Minho Hwang, Edward Cen, Ajay Kumar Tanwani, Ashwin Balakrishna, Brijen Thananjeyan, Jeffrey Ichnowski, Nawid Jamali, et al. Deep imitation learning of sequential fabric smoothing from an algorithmic supervisor. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9651–9658. IEEE, 2020. 3

[33] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575. IEEE, 2021. 1, 3

[34] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 43–52, 2021. 1, 2, 4

[35] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. In *ICRA*, 2022. 2

[36] Priya Sundaresan, Jennifer Grannen, Brijen Thananjeyan, Ashwin Balakrishna, Michael Laskey, Kevin Stone, Joseph E Gonzalez, and Ken Goldberg. Learning rope manipulation policies using dense object descriptors trained on synthetic depth data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9411–9418. IEEE, 2020. 2

[37] Neha Sunil, Shaoxiong Wang, Yu She, Edward Adelson, and Alberto Rodriguez Garcia. Visuotactile affordances for cloth manipulation with local control. In *Proceedings of The 6th Conference on Robot Learning*, pages 1596–1606. PMLR, 2023. 3

[38] Sashank Tirumala, Thomas Weng, Daniel Seita, Oliver Kroemer, Zeynep Temel, and David Held. Learning to singulate layers of cloth using tactile feedback. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7773–7780, 2022. 3

[39] Matthias Vestner, Zorah Lähner, Amit Boyarski, Or Litany, Ron Slossberg, Tal Remez, Emanuele Rodola, Alex Bronstein, Michael Bronstein, Ron Kimmel, et al. Efficient deformable shape correspondence via kernel matching. In *2017 international conference on 3D vision (3DV)*, pages 517–526. IEEE, 2017. 5

[40] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. *European conference on computer vision (ECCV 2022)*, 2022. 2

[41] Yufei Wang, Zhanyi Sun, Zackory Erickson, and David Held. One policy to dress them all: Learning to dress people with diverse poses and garments. In *Robotics: Science and Systems (RSS)*, 2023. 1, 3

[42] Thomas Weng, Sujay Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *Conference on Robot Learning*, 2021. 3

[43] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. VAT-mart: Learning visual action trajectory proposals for manipulating 3d ARTiculated objects. In *International Conference on Learning Representations*, 2022. 2

[44] Ruihai Wu, Kai Cheng, Yan Shen, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 8

[45] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 6

[46] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. In *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020. 1

[47] Chi Xu, Lakshmi Narasimhan Govindarajan, Yu Zhang, and Li Cheng. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *International Journal of Computer Vision*, 123:454–478, 2017. 2

[48] Ran Xu, Yan Shen, Xiaoqi Li, Ruihai Wu, and Hao Dong. Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation. *arXiv preprint arXiv:2403.08355*, 2024. 2

[49] Zhenjia Xu, Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Dextairity: Deformable manipulation can be a breeze. *RSS*, 2022. 6

[50] Han Xue, Yutong Li, Wenqiang Xu, Huanyu Li, Dongzhe Zheng, and Cewu Lu. Unifolding: Towards sample-efficient, scalable, and generalizable robotic garment folding. In *7th Annual Conference on Robot Learning*, 2023. 1, 3, 6

[51] 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. 1

[52] Zhengrong Xue, Zhecheng Yuan, Jiashun Wang, Xueqian Wang, Yang Gao, and Huazhe Xu. Useek: Unsupervised se (3)-equivariant 3d keypoints for generalizable manipulation. *ICRA*, 2023. 3

[53] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6496–6503. IEEE, 2022. 2, 3

[54] Fan Zhang and Yiannis Demiris. Learning grasping points for garment manipulation in robot-assisted dressing. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9114–9120. IEEE, 2020. 3

[55] Fan Zhang and Yiannis Demiris. Learning garment manipulation policies toward robot-assisted dressing. *Science Robotics*, 7(65):eabm6010, 2022. 3

[56] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *International Conference on Learning Representations (ICLR)*, 2023. 2

[57] Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, Haonan Chen, Cewu Lu, and Lin Shao. Clothesnet: An information-rich 3d garment model repository with simulated clothes environment. *ICCV*, 2023. 1, 2

[58] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 512–530. Springer, 2020. 1