2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Cloth Funnels: Canonicalized-Alignment for Multi-Purpose Garment Manipulation

Alper Canberk [1], Cheng Chi [1], Huy Ha [1],
Benjamin Burchfiel [2], Eric Cousineau [2], Siyuan Feng [2] and Shuran Song [1]
clothfunnels.cs.columbia.edu

***Abstract*— Automating garment manipulation is challenging due to extremely high variability in object configurations. To reduce this intrinsic variation, we introduce the task of "canonicalized-alignment" that simplifies downstream applications by reducing the possible garment configurations. This task can be considered as "cloth state funnel" that manipulates arbitrarily configured clothing items into a predefined deformable configuration (i.e. canonicalization) at an appropriate rigid pose (i.e. alignment). In the end, the cloth items will result in a compact set of structured and highly visible configurations – which are desirable for downstream manipulation skills. To enable this task, we propose a novel canonicalized-alignment objective that effectively guides learning to avoid adverse local minima during learning. Using this objective, we learn a multi-arm, multi-primitive policy that strategically chooses between dynamic flings and quasi-static pick and place actions to achieve efficient canonicalized-alignment. We evaluate this approach on a real-world ironing and folding system that relies on this learned policy as the common first step. Empirically, we demonstrate that our task-agnostic canonicalized-alignment can enable even simple manually-designed policies to work well where they were previously inadequate, thus bridging the gap between automated non-deformable manufacturing and deformable manipulation.**

I. Introduction

Why has garment manipulation proved more difficult to automate than more typical rigid and articulated objects? We argue that two key factors are severe self occlusion, which is present in the large set of possible crumpled states, and the infinite degrees of freedom inherent to clothing. As a result, it is impractical to manually define manipulation policies that achieve reliable manipulation — a cornerstone of current automated non-deformable manufacturing pipelines. In this work, we explore bridging the gap between existing approaches to automation and the challenging domain of clothing. We show that when a robot first manipulates arbitrarily configured clothing items into a predefined configuration (*i.e.* canonicalization) at an appropriate pose (*i.e.* alignment), downstream manipulation skills work significantly more reliably.

Recently, real-world cloth manipulation has received significant attention. Some of the earliest cloth manipulation work explored manually designed heuristics which worked well for specific clothing types, configurations, and tasks, such as cloth unfolding [1–4], smoothing [5, 6], folding [2, 7–9], but their strong assumptions initial states, fiducial markers, specialized tools, or cloth type/shape do not generalize. More recently, learning-based approaches have shown success in more general cloth manipulation behavior. One line of work has explored supervised-learning from human demonstrations for smoothing [10] and folding [7], but those methods required costly human demonstrations/annotations. Another recent line of work employs fully self-supervised learning and has shown success in learning to unfold [11] (but doesn't generalize to other tasks) and in tackling visual goal-conditioned manipulation of a single square cloth instance [12].

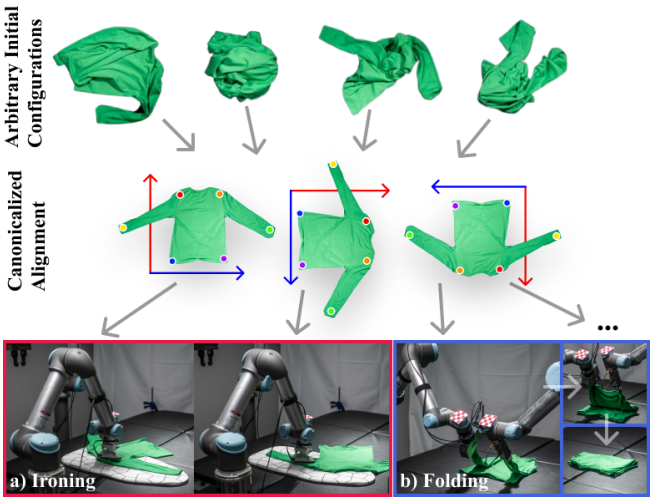

Fig. 1. Canonicalized-Alignment *funnels* the large space of possible cloth configurations into a much smaller and better structured set of highly-visible states that greatly simplifies downstream tasks such as ironing or folding.

Instead of learning arbitrary monolithic cloth manipulation tasks, we hypothesize that it is more efficient to learn a robust task-agnostic canonicalization and alignment policy from which other task-specific manipulation skills may be chained. This is because such a policy *funnels* unstructured and self-occluded cloth configurations into structured states with clearly visible key points (Fig. 1, middle), enabling even the simplest heuristics to achieve a high success rate on downstream tasks.

To this end, our *primary contribution* is the introduction of the "canonicalized-alignment", a garment manipulation task which serves as a cloth funnel for reducing general-purpose garment manipulation complexity. The goal of this task is to manipulate a garment into a canonical shape (defined by its category) and align it with a particular 2D translation and rotation. We tackle this task through several technical contributions.

- We propose a learned multi-arm, multi-primitive manipulation policy that strategically chooses between dynamic flings and quasi-static pick&place actions to efficiently and precisely transform the garment into its

[1] Columbia University
[2] Toyota Research Institute

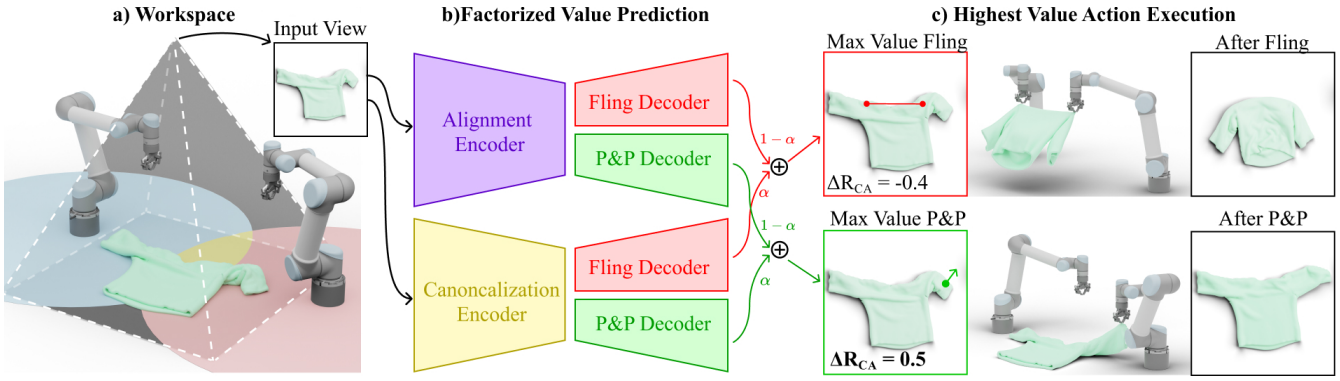

Fig. 2. **Approach Overview.** a) A batch of scaled and rotated observations are created from a top-down RGB image of the workspace and then concatenated with a scale-invariant coordinate map. b) The batch of inputs is fed through the factorized network architecture, producing a batch of rotated and scaled value maps for each primitive. c) All primitive batches are concatenated and the maximum value pixel parameterizes the action to be executed.

canonicalized and aligned configuration.

- To train the policy, we proposed a novel factorized reward function that avoids adverse local minima which plague the generic goal-reaching formulations by decoupling deformable shape and rigid pose.
- We evaluate our approach in multiple downstream garment manipulation tasks in the real-world on a physical robot, including folding and ironing.

Our experiments show that incorporation of canonicalized-alignment significantly reduces the complexity of downstream applications, suggesting that robust canonicalized-alignment provides a practical step forward toward multi-purpose garment manipulation from arbitrary states for diverse tasks.

II. Related Work

Heuristic-based cloth manipulation. Heuristic-based manipulation pipelines – where action selection and planning is manually designed – can produce impressive results. However, the generality and robustness of these approaches is limited due to strong assumptions regarding pre-canonicalized initial state [9, 13], fiducial markers [14], specialized tools [8], and cloth type and shape [1, 2, 4–6, 15–18].

Learning-based cloth unfolding. Learning-based methods can self-discover the best policies for a distribution of cloths using real-world self-supervision [11, 19] or simulator states [20, 21]. While these approaches have been successfully applied to cloth unfolding [11, 19] or canonicalization [20], they do not consider canonicalized-alignment. This limits their applicability since heuristic-based pipelines cope poorly with unmet cloth assumptions or kinematic constraints.

Goal-conditioned cloth manipulation. Towards generic goal-conditioned cloth manipulation, prior works have investigated reinforcement learning [22–25], real-world self-supervised learning [12] and imitation learning [26]. However, these methods often struggle to bridge the sim2real gap [24], generalize across cloth instances [12, 26, 27] or generalize between garment types [23, 25, 28]. Furthermore, all goal-conditioned works do not address how goal vertices/key points/images can be obtained for a completely novel cloth instance. Instead, our proposed approach can accommodate different garment categories and generalize to a variety of novel real-world garment instances from simulation training.

III. Method

A. A Multi-Purpose Garment Manipulation Pipeline

We propose a factorized approach to multi-purpose garment manipulation from arbitrary states that decomposes the process into two steps. First, the robot executes a learned *task-agnostic* canonicalized-alignment policy, which leaves the garment in a known configuration predefined for the clothing category at a specified 2D rotation and translation. Second, the robot executes a *task-specific* keypoint based policy, which could be as simple as a manually-designed heuristic. This approach confers three primary benefits:

- **Arbitrary initial configuration**: Canonicalization *funnels* the large space of possible cloth configurations into a narrow distribution of highly-structured fully-observable configurations from which downstream policies can more easily operate.
- **Downstream task-awareness**: Flexible goal-conditioned alignment allows the canonicalized cloths to be placed at specified positions and orientations that are kinematically appropriate for particular downstream tasks.
- **Clothing category generalization**: A keypoint-based cloth representation effectively reduces the observation space from having to represent the infinite DoF down to a few meaningful keypoints. Further, cloths are always in a known canonicalized configuration. These two properties combined not only simplifies learning downstream task-specific manipulation policies, but also makes it possible to engineer heuristics that work reliably for a clothing category.

Next, we will discuss how the canonicalized-alignment task is formulated (Sec. III-B), learned (Sec. III-C), and implemented alongside the several downstream task policies (Sec. III-D).

B. The Canonicalization & Alignment Task

Problem Formulation. Given a clothing item in some clothing category in an arbitrary initial configuration, the goal of canonicalization is to reach the human-defined standard deformable configuration for that clothing category, such as a T-shaped configuration for shirts and the upside-down V-shaped configuration for pants. Note that this only accounts for the garment's shape, but not its pose in the workspace. This means canonicalization alone can't ensure that downstream tasks are kinematically feasible. To address this shortcoming, the goal of "canonicalized-alignment" is to reach canonicalization at a specific planar position and rotation in the workspace.

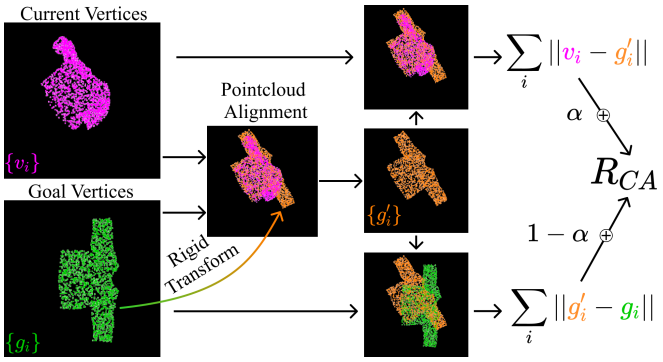

Fig. 3. **Reward Computation.** From the goal configuration $\mathbf{g}$ (green) and current configuration $\mathbf{v}$ (magenta), we compute a best-alignment configuration $\mathbf{g}'$ (orange) based on the ground-truth vertex correspondence between the two configurations of the same cloth mesh. Then, the average vertex distance between $\mathbf{g}'$ and $\mathbf{g}$ (where only rigid transforms matter) gives the alignment reward R_{A}, while that between $\mathbf{g}$ and $\mathbf{v}$ (where only deformation matters) gives the canonicalization reward R_{C}. Our factorized reward R_{CA} is a weighed sum between R_{A} and R_{C}.

Naive Reward Formulation. Given a simulated cloth instance with N vertices, let $\mathbf{v}=\{v_i\}_{i\in[N]}$ denote the current configuration of the cloth (Fig. 3, magenta), where each $v_i \in \mathbb{R}^3$ is the position of the ith cloth vertex. Given a goal configuration $\mathbf{g}=\{g_i\}_{i\in[N]}$ and the vertex correspondence between $\mathbf{g}$ and $\mathbf{v}$ using the cloth mesh in simulation, the average per-vertex distance between $\mathbf{g}$ and $\mathbf{v}$ gives a generic goal-conditioned cloth manipulation cost. In the specific case where $\mathbf{g}$ is a canonicalized configuration of the cloth at the goal position and rotation (Fig. 3, green), we have the following straightforward canonicalized-alignment reward

$$R_{\text{Unf}} = -||\mathbf{g}-\mathbf{v}||_2 \tag{1}$$

Clearly, R_{Unf} is consistent, in that a policy which achieves $R_{\text{Unf}}=0$ achieves perfect canonicalized-alignment. However, this formulation has two primary downsides:

1) **Entangled supervision.** When R_{Unf} is low, it can be difficult for the policy to tell whether it should make a planar transformation of the cloth configuration (such as shifting entire cloth to the right) or a deformable adjustment (such as flipping a shirt's sleeve outwards).
2) **Over-emphasis on cloth pose.** Actions that shift the cloth result in sharp and large changes in R_{Unf}, while actions of smaller magnitudes become insignificant. Since such small adjustment actions are required to bring a poorly canonicalized cloth to a well-canonicalized one, R_{Unf} fails to put enough emphasis on the canonicalization subtask, and leads to a problematic local minima in policy learning.

Factorized Reward Formulation. To alleviate these shortcomings, we propose a reward factorization, that expresses the canonicalization R_{C} and alignment R_{A} aspects of the task separately:

$$R_{\text{CA}} = \alpha R_{\text{C}} + (1-\alpha) R_{\text{A}} \tag{2}$$

where $\alpha \in (0,1)$ is a hyperparameter. With this factorization, we can provide separate supervision R_{C} and R_{A} during training, while acting with respect to R_{CA} during data-collection and inference. This helps the policy distinguish how actions separately affect the cloth shape or planar pose. Further, with a tunable α, we can emphasize R_{C} more than R_{A}, which empirically leads to better canonicalization results (Tab. III).

To factorize the reward, we propose to compute a transform T, which transforms $\mathbf{g}$ into a best-aligned goal configuration $\mathbf{g}'$ (Fig. 3, orange). Given such a $\mathbf{g}'$, its distance to $\mathbf{v}$ accounts only for their deformable shape mismatch, which serves as the canonicalization reward,

$$R_{\text{C}} = -||\mathbf{v}-\mathbf{g}'||_2 \tag{3}$$

Meanwhile, by T's definition, the distance between $\mathbf{g}'$ and $\mathbf{g}$ accounts only for the mismatch in planar position and rotation, which serves as the alignment reward,

$$R_{\text{A}} = -||\mathbf{g}'-\mathbf{g}||_2. \tag{4}$$

Factorization Implementation. To find $\mathbf{g}'$, we have observed that naively minimizing the average per-vertex distance between $\mathbf{g}$ and $\mathbf{v}$ is extremely sensitive to outliers, so does not give us the best alignment. Such outliers arise due to mismatches in $\mathbf{g}$'s and $\mathbf{v}$'s deformable shapes where small protrusions with large offsets (*e.g.* a shirt's arm folded inwards) could significantly shift the minimum distance configuration. To filter out such outliers, we optimize the transform T which minimizes this distance for only a subset of points, where point i is included if $||g_i - v_i||_2 \leq \tau$ for some scalar threshold τ then apply T to $\mathbf{g}$ to get $\mathbf{g}'$. We repeat this minimization and filter procedure in iterations, using the previous iteration's $\mathbf{g}'$ as the current $\mathbf{g}$, until convergence.

In our experiments, we observed that $\alpha=0.6$ and $\tau=0.3$ performs best. To account for different cloth sizes, we normalize all R_{C}, R_{A}, R_{Unf}, and τ by the geometric mean of the cloth's height and width in a canonicalized configuration. Since most garments are mirror-symmetric, we select the highest reward from either the goal configuration or its mirror-flip in the goal's local frame.

C. Multi-Primitive Spatial Action Maps Policy

Coarse-grain dynamic multi-arm flings can efficiently unfold and align garments from crumpled states [11], but are insufficient for the fine-grained adjustments required to achieve canonicalization. To overcome this challenge, we propose a multi-arm, multi-primitive system that combines quasi-static and dynamic actions, which enables both efficient *and* fine-grained manipulation. To unify the primitive parametrizations and easily enforce constraints, we use a spatial action maps policy.

Spatial Action Maps is a convolutional neural network [29] (CNN) policy for learning value maps [30] where actions are defined on a pixel grid. Through its simple and effective exploitation of translational, rotational, and scale equivariances, spatial action maps is a popular framework for learning robotic policies [11, 19, 31].

We extend FlingBot [11]'s spatial action maps approach as follows. Given a $H \times W \times 3$ top-down view of the workspace (Fig. 2a), we rotate and scale it to form a stack of transformed observations of shape $K \times H \times W \times 3$. To help the network reason about the cloth's alignment, we concatenate a $K \times H \times W \times 2$ scale-invariant, normalized (between -1 and 1) positional encoding to the transformed observation stack. To enable multiple primitives, we propose a factorized network architecture (Fig. 2b) with two encoders, one for each task's reward (R_{C}, R_{A}), where each encoder has two decoder heads, one for each primitive. The encoders take in the transformed observation stack, and the decoders output value maps, one for each reward-primitive pair. The value maps are combined using (2), and the highest value action (over all action parameters and primitives) is chosen (Fig. 2c).

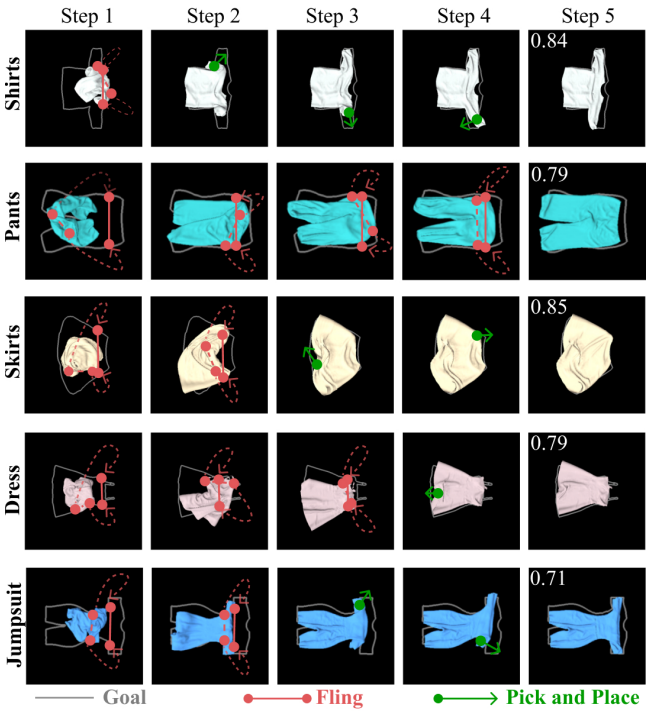

Fig. 4. **Canonicalized-Alignment of Multiple Categories.** In each row, we demonstrate a sequence of the first 5 actions taken by the model corresponding to a clothing category in simulation.

System Implementation. As our spatial action map backbone, we use the factorized network architecture (Fig. 2b), the encoders and decoders of which are from a 6-layer UNet [32]. In our experiments, we consider two action primitives, quasi-static pick&place and dynamic flings. We use $(H,W) = (128,128)$ and a decaying ϵ-greedy for exploration of action primitives (*i.e.* fling v.s. pick&place) and action parameters. By constraining the observation's transforms to 16 rotations spanning 360° and scales in $\{0.75,1.0,1.5,2.0,2.5,3.0\}$ (giving $K = 96$), we can ensure that arms neither collide nor crossover each other. Our value networks' predictions are supervised using the delta-reward values – which is the difference in R_{CA} before and after an action is taken – using the MSE loss and the Adam [33] optimizer with a learning rate of $1e-4$.

D. Keypoint-based Task Heuristics

Compared to learning-based approaches, heuristics are highly interpretable and thus simple to define. Here, we demonstrate that it's possible to use heuristics for shirt ironing with no keypoints and folding with a small set of keypoints.

Keypoint Detection. We collect 200 cloth configurations from simulation with coverage at least 60% and trained a DeepLabv3 [34] detector for each garment class. Using a random 80/20 training/evaluation split, we observed that this detector generalizes well to novel garment instances with average error of 5/128 pixels. Setting up a keypoint detector model for a new clothing category takes roughly 1 hour. After detection, each keypoint is depth-projected into 3D points and transformed into the workspace frame of reference. By representing cloths as a set keypoints, we sidestep their infinite DoF by using a few meaningful keypoints as the representation, which makes it simpler to define heuristics over them. For instance, long sleeve shirts have six keypoints for two sleeves, shoulders, and waists.

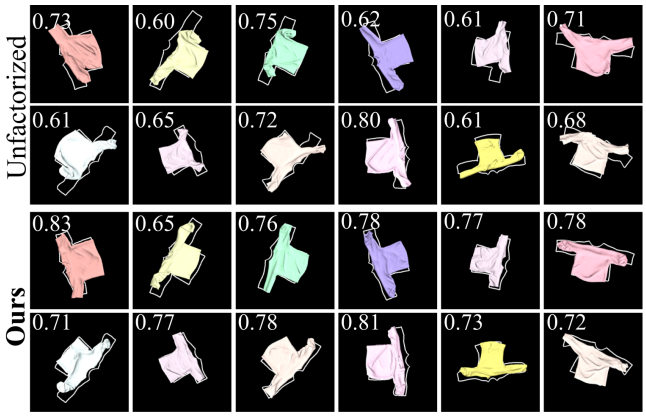

Fig. 5. **Reward Comparisons.** We qualitatively compare the final configuration achieved by policies trained with factorized and the unfactorized reward formulation. For the qualitative comparisons, the IoU of the final frame of various rollouts are shown in top-left corner of each square.

Ironing Heuristic. For garment manipulation pipelines, specialized tools are placed at a fixed location in the workspace. For ironing, the extra tools involved are the ironing board and the arm holding the iron (Fig. 8 right). Given a well canonicalized and aligned shirt, an open-loop ironing primitive where the end effector moves from one end of the ironing board and back without any perception can be sufficient. In our setup, we use two transforms such that the left and right side of the shirt is on the ironing board respectively.

Folding Heuristic. First, the sleeves are folded towards the waist using a dual-arm pick and place action. Here, the pick point is the sleeve keypoint, while the place point is the quarter and three-quarter point along the waist line (computed from the waist keypoints). Since not all shirt arms are long enough to reach the waist points, the place points are constrained to be an arm's length distance away from the shoulder keypoints. The arm length can be computed from keypoints as the minimum distance between the sleeve and the shoulder keypoints over the left and right arms. In the second step, with the arms folded in, the shoulder keypoints are picked and placed at the waist keypoints (Fig. 8 left).

IV. Results

In simulation, we conduct ablation studies of reward formulation (Sec. IV-C) and action primitives (Sec. IV-D). Next, we demonstrated our approach on five garment categories from Cloth3D [35] (Sec. IV-E) and a folding downstream task (Sec. IV-F). **In the real world**, we include primitive and reward comparisons for the long-sleeve shirt category on canonicalized-alignment (Tab. IV, Fig. 7), folding and ironing.

A. Metrics

After running each policy for 10 steps, we evaluate the final R_{Unf} (1), R_{A} (3), and R_{C} (4) on the last step. These rewards measure distance based on effectively ground-truth vertex correspondence, which means that they can't readily be computed on real world garments. To address these shortcomings, we also evaluate the intersection over union (IoU) and percent coverage from the current cloth binary image mask and the goal cloth binary image mask. In all tables, $\downarrow$ indicates that lower is better while $\uparrow$ indicates that larger values are preferred.

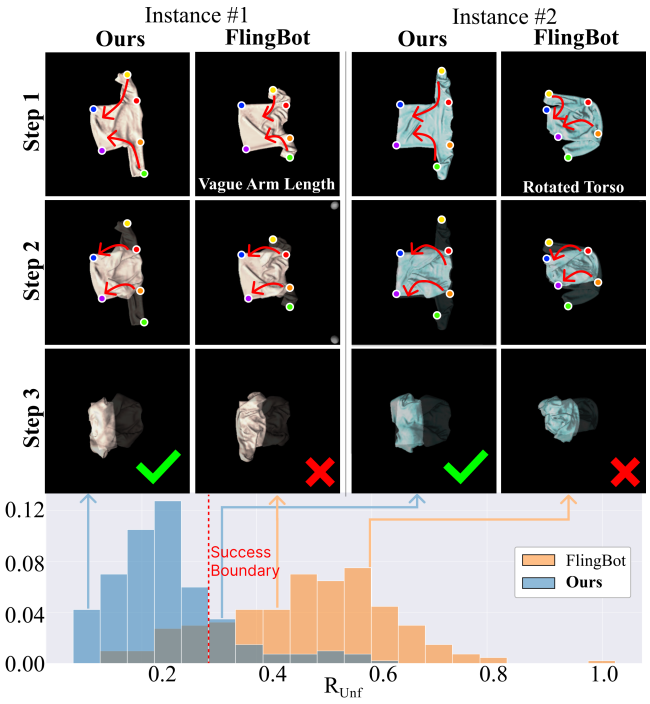

Fig. 6. **Simulation Folding Qualitative Comparison and Frequency Histgoram.** We compare our folding heuristic on our method's canonicalization results vs. FlingBot's unfolding results. We demonstrate that not all high-coverage configurations ensure folding success.

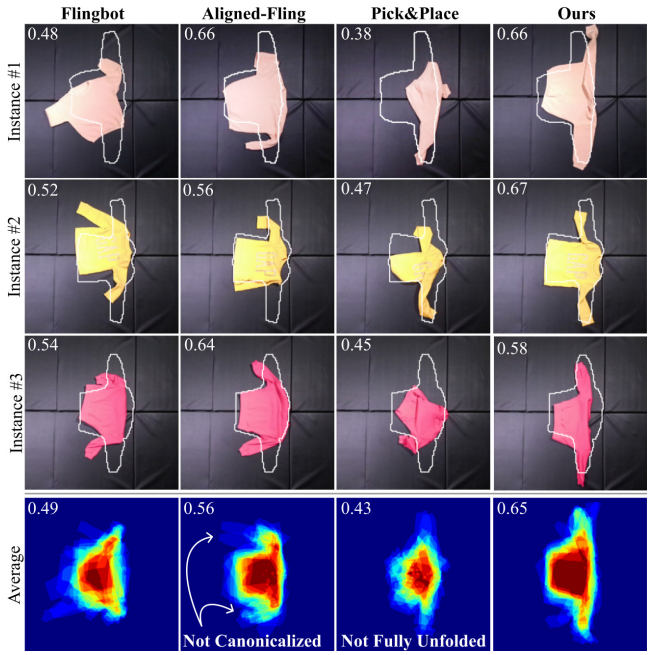

Fig. 7. **Real-world Canonicalized-Alignment.** High-coverage configurations achieved by Flingbot aren't always aligned, which is improved with our Aligned-Fling. However, using only coarse-grained Aligned-Fling fails to fully canonicalize the shirt, and using only fine-grained Pick & Place fails to fully-unfold the shirt as can be seen by the average of the final cloth masks (bottom row).

B. Task Generation

Our task datasets contain randomized initial configurations of a filtered[1] subset of meshes from Cloth3D [35], whose configurations are generated as follows:

1) **Hard Tasks** have low coverage and severe self-occlusion. They are generated by randomly rotating the cloth, picking a random point on the cloth, dropping it from a random height in [0.5,1.5]m, and then translating the cloth by a random distance in [0.0,0.2]m.
2) **Easy Tasks** have high coverage to test policies' abilities to perform small adjustments crucial to canonicalization. They are generated by starting with the canonicalized configuration, and dragging a random point on the cloth by an angle uniformly sampled from [0,360] degrees by a distance uniformly sampled in [0.5,1]m.

Each garment category has 2000 training and 400 testing tasks with unseen meshes, with a 75-25 and 50-50 split between hard and easy tasks respectively.

C. Reward Ablation

We compare the canonicalized-alignment performance between the unfactorized (1) and our factorized reward formulation (2) on the long sleeve category. We observe that our approach consistently does best on all metrics (Tab. III), reflected in qualitatively more consistent canonicalized-alignment (Fig. 5). We hypothesize that the R_{Unf} baseline struggles to canonicalize properly due to faint supervision on small deformable adjustments. Meanwhile, our approach can emphasize canonicalization more with $\alpha = 0.6$.

[1]Since CLOTH3D meshes are automatically generated, we manually filter unrealistic mesh examples (*e.g.* arms as wide as cloth body), and we ensured all cloth meshes are shorter than 0.7m.

D. Effectiveness of Combined Primitives

While high-velocity dynamic actions enable efficient unfolding, precise quasi-static actions are necessary for fine-grained adjustments involved in canonicalization. We compare two single primitive systems, only Aligned-Fling (Flingbot [11]'s fling with fling direction and location specified by the target alignment) and only Pick & Place (P&P), with our combined primitive system on canonicalized-alignment of long sleeve shirts.

On easy tasks in simulation (Tab. I, top), P&P and our combined primitive system perform similarly, confirming that quasi-static actions alone are effective for small adjustments. In contrast, the fling-only system performs poorly because the imprecise flinging actions are ill-suited for cases where fine-grained manipulation is required (*e.g.* Fig. 2).

For hard tasks (Tab. I bottom in simulation, Tab. IV in real), we observe that both single primitive baselines perform poorly. Notably, Aligned-Fling achieves a similar performance between easy and hard tasks in simulation, suggesting that the effect of coarse grain high-velocity actions can be effective for unfolding but is not versatile enough for canonicalization. Meanwhile, the P&P baseline performed significantly worse on hard tasks in simulation, confirming the findings reported in FlingBot [11]. Our combined primitive system achieves the best performance, demonstrating the synergy between coarse-grain high-velocity flings and fine-grain quasi-static actions for canonicalized-alignment.

Due to imperfect cloth simulators, deformable dynamics such as arms twisting (instance #2, last column in Fig. 7) and stretching (instance #3, last column in Fig. 7) are never observed in simulation. Despite this sim2real gap, our real world qualitative results (Fig. 7) confirmed our simulation

findings. Specifically, we found that on average, our combined primitive approach achieved an IoU of 0.65, while other single primitive baselines/ablations reached 0.56 or less.

E. Canonicalized-Alignment on Garment Category

Using the same learning and system configuration, we trained new models on 4 other garment categories, one model per garment category. From the quantitative evaluation in Tab. II, our approach achieves around 0.7 IoU for all categories, which demonstrates the generality of our problem formulation with respect to garment categories. Further, from Fig. 4, our policy has learned that while flinging is crucial for quickly unfolding crumpled garments, a few pick&places are usually required to achieve good canonicalization.

TABLE I

PRIMITIVE ABLATION IN SIMULATION

Task	Primitives	$R_{Unf}\downarrow$	$R_A\downarrow$	$R_C\downarrow$	IoU ↑	Cov.↑
Easy	Aligned-Fling	0.100	0.058	0.079	0.649	0.821
	P&P	0.077	0.068	**0.037**	**0.734**	**0.928**
	Aligned-Fling+P&P	**0.075**	**0.051**	0.044	0.731	0.924
Hard	Aligned-Fling	0.100	0.058	0.079	0.644	0.812
	P&P	0.136	0.111	0.077	0.601	0.792
	Aligned-Fling+P&P	**0.075**	**0.051**	**0.052**	**0.728**	**0.887**

TABLE II

EVALUATION ON MULTIPLE CATEGORIES ON HARD TASKS

Category	$R_{Unf}\downarrow$	$R_A\downarrow$	$R_C\downarrow$	IoU ↑	Cov.↑
Shirt	0.075	0.051	0.052	0.728	0.887
Pants	0.098	0.077	0.053	0.708	0.892
Skirt	0.159	0.128	0.122	0.680	0.837
Dress	0.149	0.106	0.100	0.714	0.878
Jumpsuit	0.099	0.072	0.060	0.648	0.817

TABLE III

REWARD ABLATION ON HARD TASKS

Metric	$R_{Unf}\downarrow$	$R_A\downarrow$	$R_C\downarrow$	IoU ↑	Cov.↑
R_{Unf}	0.093	0.069	0.064	0.684	0.879
R_{CA} ($\alpha=0.6$)	**0.075**	**0.051**	**0.052**	**0.728**	**0.887**

F. Application in Downstream Cloth Manipulation

A primary motivation for this work is the improvement of downstream tasks; we hypothesize that effective canonicalized-alignment will significantly reduce the complexity of subsequent manipulation skills. To this end, we study ironing and folding (Sec. III-D). For folding, we measured the final R_{Unf} achieved by each approach when the goal configuration is the ground-truth folded configuration at a specified alignment. We also include a folding success rate which is a thresholded R_{Unf}, where the boundary is chosen by qualitatively.

Canonicalized-Alignment Improves Downstream Tasks. From Tab. V, we observe that manually-designed folding heuristic completely fails at the task if starting from random initial configurations. While success rate improves with FlingBot [11]'s unfolded configurations, achieving high-coverage configurations (the goal of unfolding) does not always give structured, high-visibility configurations required by heuristics (Fig. 6), so FlingBot [11] still performs poorly. With canonicalized-alignment, policies trained with R_{Unf} and R_{CA} achieve success rates of 84.9 and 87.8 respectively, demonstrating the importance of canonicalized-alignment for reducing downstream task complexity, such that even simple manually designed task heuristics can work well.

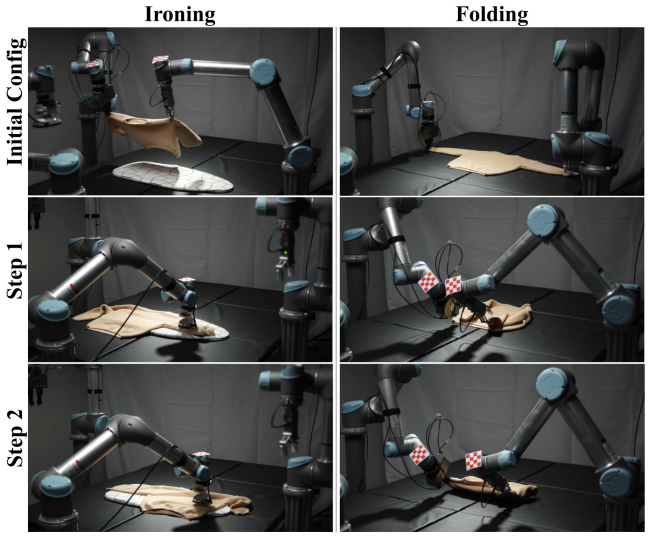

Fig. 8. **Real-world Ironing & Folding.** Reliable canonicalized-alignment not only gives high-visibility starting configurations, which reduces complexity for downstream tasks like folding (step 1), but can also be called multiple times with different target alignments, which is useful in ironing (step 2 & 3). More video results on the project website.

TABLE IV

REAL-WORLD CANONICALIZED-ALIGNMENT

Approach	IoU ↑	Cov.↑
FlingBot [11]	0.489	0.709
Aligned-Fling	0.558	0.735
P&P	0.433	0.507
Aligned-Fling+P&P	**0.648**	**0.806**

TABLE V

SIMULATION FOLDING

Approach	Success ↑	$R_{Unf}\downarrow$
Random	2.1	0.520
FlingBot [11]	19.6	0.486
R_{Unf}	84.9	**0.246**
R_{CA} (Ours)	**87.8**	0.253

V. CONCLUSION

We introduce the task of canonicalized-alignment, a universal first step for multi-purpose garment manipulation pipelines. By funneling a diverse array of crumpled clothing into a small set of high-visibility configurations, this task addresses much of the complexity associated with cloth's infinite DOF state space and severe self-occlusion, and thus simplifies downstream tasks such as ironing and folding. Due to imperfect cloth simulators, we hypothesize that canonicalized-alignment performance can be improved if a real-world supervision signal could be derived for enabling real-world finetuning, and we believe this is an interesting direction for future work.

Acknowledgements. This work was supported in part by the Toyota Research Institute, NSF Award #2143601, #2037101, and #2132519. We would like to thank Google for the UR5 robot hardware. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] Marco Cusumano-Towner, Arjun Singh, Stephen Miller, James F O'Brien, and Pieter Abbeel. Bringing clothing into desired configurations with limited perception. In *2011 IEEE international conference on robotics and automation*, pages 3893–3900. IEEE, 2011.

[2] Jeremy Maitin-Shepard, Marco Cusumano-Towner, Jinna Lei, and Pieter Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315. IEEE, 2010.

[3] Dimitra Triantafyllou, Ioannis Mariolis, Andreas Kargakos, Sotiris Malassiotis, and Nikos Aspragathos. A geometric approach to robotic unfolding of garments. *Robotics and Autonomous Systems*, 75:233–243, 2016.

[4] Fumiaki Osawa, Hiroaki Seki, and Yoshitsugu Kamiya. Unfolding of massive laundry and classification types by dual manipulator. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 11(5):457–463, 2007.

[5] Li Sun, G. Aragon-Camarasa, W. Cockshott, Simon Rogers, and J. Siebert. A heuristic-based approach for flattening wrinkled clothes. In *TAROS*, 2013.

[6] B. Willimon, S. Birchfield, and I. Walker. Model for unfolding laundry using interactive perception. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4871–4876, 2011. doi: 10.1109/IROS.2011.6095066.

[7] Peng Zhou, Omar Zahra, Anqing Duan, Shengzeng Huo, Zeyu Wu, and David Navarro-Alarcon. Learning cloth folding tasks with refined flow based spatio-temporal graphs. *arXiv preprint arXiv:2110.08620*, 2021.

[8] Fumiaki Osawa, Hiroaki Seki, and Yoshitsugu Kamiya. Clothes folding task by tool-using robot. *Journal of Robotics and Mechatronics*, 18(5):618–625, 2006.

[9] Stephen Miller, Mario Fritz, Trevor Darrell, and Pieter Abbeel. Parametrized shape models for clothing. In *2011 IEEE International Conference on Robotics and Automation*, pages 4861–4868. IEEE, 2011.

[10] 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)*. IEEE, 2022.

[11] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robotic Learning (CoRL)*, 2021.

[12] Robert Lee, Daniel Ward, Akansel Cosgun, Vibhavari Dasagi, Peter Corke, and Jurgen Leitner. Learning arbitrary-goal fabric folding with one hour of real robot experience, 2020.

[13] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *ICLR*, 2019.

[14] Christian Bersch, Benjamin Pitzer, and Sören Kammel. Bimanual robotic cloth manipulation for laundry folding. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1413–1419. IEEE, 2011.

[15] Andreas Doumanoglou, Jan Stria, Georgia Peleka, Ioannis Mariolis, Vladimir Petrik, Andreas Kargakos, Libor Wagner, Václav Hlaváč, Tae-Kyun Kim, and Sotiris Malassiotis. Folding clothes autonomously: A complete pipeline. *IEEE Transactions on Robotics*, 32 (6):1461–1478, 2016.

[16] Kenta Tanaka, Yusuke Kamotani, and Yasuyoshi Yokokohji. Origami folding by a robotic hand. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2540–2547. IEEE, 2007.

[17] Devin J Balkcom and Matthew T Mason. Robotic origami folding. *The International Journal of Robotics Research*, 27(5):613–627, 2008.

[18] Jan Stria, Daniel Průša, Václav Hlaváč, Libor Wagner, Vladimír Petrík, Pavel Krsek, and Vladimír Smutný. Garment perception and its folding using a dual-arm robot. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 61–67, 2014. doi: 10.1109/IROS.2014.6942541.

[19] Zhenjia Xu, Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Dextairity: Deformable manipulation can be a breeze. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[20] Zixuan Huang, Xingyu Lin, and David Held. Mesh-based dynamics with occlusion reasoning for cloth manipulation. *arXiv preprint arXiv:2206.02881*, 2022.

[21] Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Conference on Robot Learning*, pages 256–266. PMLR, 2022.

[22] 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.

[23] Julius Hietala, David Blanco-Mulero, Gokhan Alcan, and Ville Kyrki. Closing the sim2real gap in dynamic cloth manipulation. *arXiv preprint arXiv:2109.04771*, 2021.

[24] 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.

[25] Yoshihisa Tsurumine, Yunduan Cui, Eiji Uchibe, and Takamitsu Matsubara. Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation. *Robotics and Autonomous Systems*, 112: 72–83, 2019.

[26] 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.

[27] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy.

In *Conference on Robot Learning*, pages 192–202. PMLR, 2022.

[28] Daisuke Tanaka, Solvi Arnold, and Kimitoshi Yamazaki. Emd net: An encode–manipulate–decode network for cloth manipulation. *IEEE Robotics and Automation Letters*, 3(3):1771–1778, 2018.

[29] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.

[30] Jimmy Wu, Xingyuan Sun, Andy Zeng, Shuran Song, Johnny Lee, Szymon Rusinkiewicz, and Thomas Funkhouser. Spatial action maps for mobile manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[31] Jimmy Wu, Xingyuan Sun, Andy Zeng, Shuran Song, Johnny Lee, Szymon Rusinkiewicz, and Thomas Funkhouser. Spatial action maps for mobile manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. doi: 10.48550/ARXIV.1505.04597. URL https://arxiv.org/abs/1505.04597.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arxiv 2017. *arXiv preprint arXiv:1706.05587*, 2, 2019.

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