A Continuous Robot Vision Approach for Predicting Shapes and Visually Perceived Weights of Garments

Li Duan[1], Gerardo Aragon-Camarasa[1]

***Abstract*—We present a continuous perception approach that learns geometric and physical similarities between garments by continuously observing a garment while a robot picks it up from a table. The aim is to capture and encode geometric and physical characteristics of a garment into a manifold where a decision can be carried out, such as predicting the garment's shape class and its visually perceived weight. Our approach features an early stop strategy, which means that a robot does not need to observe a full video sequence of a garment being picked up from a crumpled to a hanging state to make a prediction, taking 8 seconds in average to classify garment shapes. In our experiments, we find that our approach achieves prediction accuracies of 93% for shape classification and 98.5% for predicting weights and advances state-of-art approaches in similar robotic perception tasks by 22% for shape classification.**

***Index Terms*—Computer Vision for Automation, Deep Learning for Visual Perception, Visual Learning, AI-Enabled Robotics**

I. Introduction

PERCEPTION and manipulation of deformable objects remain an open problem in robotics. This is because garments have an infinite number of possible configurations that cannot be modelled easily via simulations [1], [2]. Specifically, the main challenges in deformable objects perception and manipulation are twofold. First, they usually have a complex initial configuration, which means they are wrinkled, crumpled or folded, and not in a known configuration state that can be used for manipulation tasks. Second, garments deform in unpredictable ways, making predictions of their deformations difficult during dexterous robotic manipulations.

Robots manipulating deformable objects without prior knowledge about their geometric and physical properties (e.g. shapes, weights or stiffness parameters) can result in robots requiring to plan actions using a complex and high-dimensional space. This causes failures in motion planning since robots are prone to fail due to minor variations in the deformable object's configurations. We propose in this paper an online continuous perception approach that equips a robot with the ability to predict garment shapes and allows a robot to stop a manipulation task if the prediction belief is above a threshold.

Manuscript received: February, 22, 2022; Revised May, 16, 2022; Accepted June, 20, 2022.

This paper was recommended for publication by Editor Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments.

[1]Li Duan and Gerardo Aragon-Camarasa are with School of Computing Science, University of Glasgow, Glasgow, United Kingdom, G12 6EU l.duan.1@research.gla.ac.uk, gerardo.aragoncamarasa@glasgow.ac.uk

Digital Object Identifier (DOI): see top of this page.

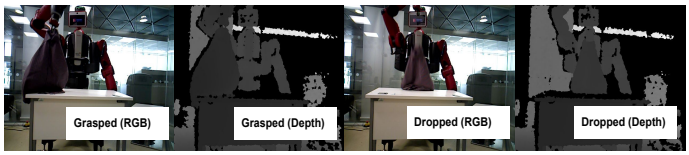

Fig. 1. An example of our experimental setup for capturing a database of garment's deformations. That is, a dual-armed robot grasps garments from a pre-defined fixed point and dropped. An Xtion camera is placed in the front of the robot to record a video sequence of RGB and depth images.

Compared to previous research focusing on wrinkles [3] and other local features [4], [5], we propose to learn the dynamic properties of garments from video sequences and allow a robotic system to recognise the shape and weight of a garment continuously. For this, we implement a Garment Similarity Network (termed GarNet in the scope of this paper) which is based on a Siamese neural network architecture that learns the physical similarity between garments to predict shapes (geometric) and visually perceived weights (physical) of unseen garments. We define a visually perceived weight in three discretised levels using an electric scale to physically weigh garments; namely, light, medium and heavy garment weights. We hypothesise *that our approach can predict shapes and discretised weights in approximately 0.1 seconds per image frame (with a size of* 256×256*) by learning geometric and physical properties and predicting the garment's shape and weights continuously during a robotic garment pick and place task.*

To test the above hypothesis, we have built a database that consists of RGBD video sequences of a robot grasping and dropping garments on a table, see Fig. 1. This database simulates a sorting scenario (e.g. [3], [4]) where a robot can sort based on shape or weights. We then train GarNet to learn garments' geometric and physical similarities based on their shape and discretised weight labels. GarNet's objective is thus to cluster garments of the same categories (shapes or discretised weights) together and pull garments of different categories apart using a triplet loss function, and these clusters are mapped into a Garment Similarity Map (GSM). To predict unseen garment shapes and weights, we introduce the concept of decision points which depend on previously mapped points in the GSM. We use these decision points to implement an early-stop strategy by fitting confidence intervals for each cluster to allow us to determine whether decision points are within a statistical significant interval around a cluster. Figure 2 shows an overview of our approach. The contributions of this paper are threefold:

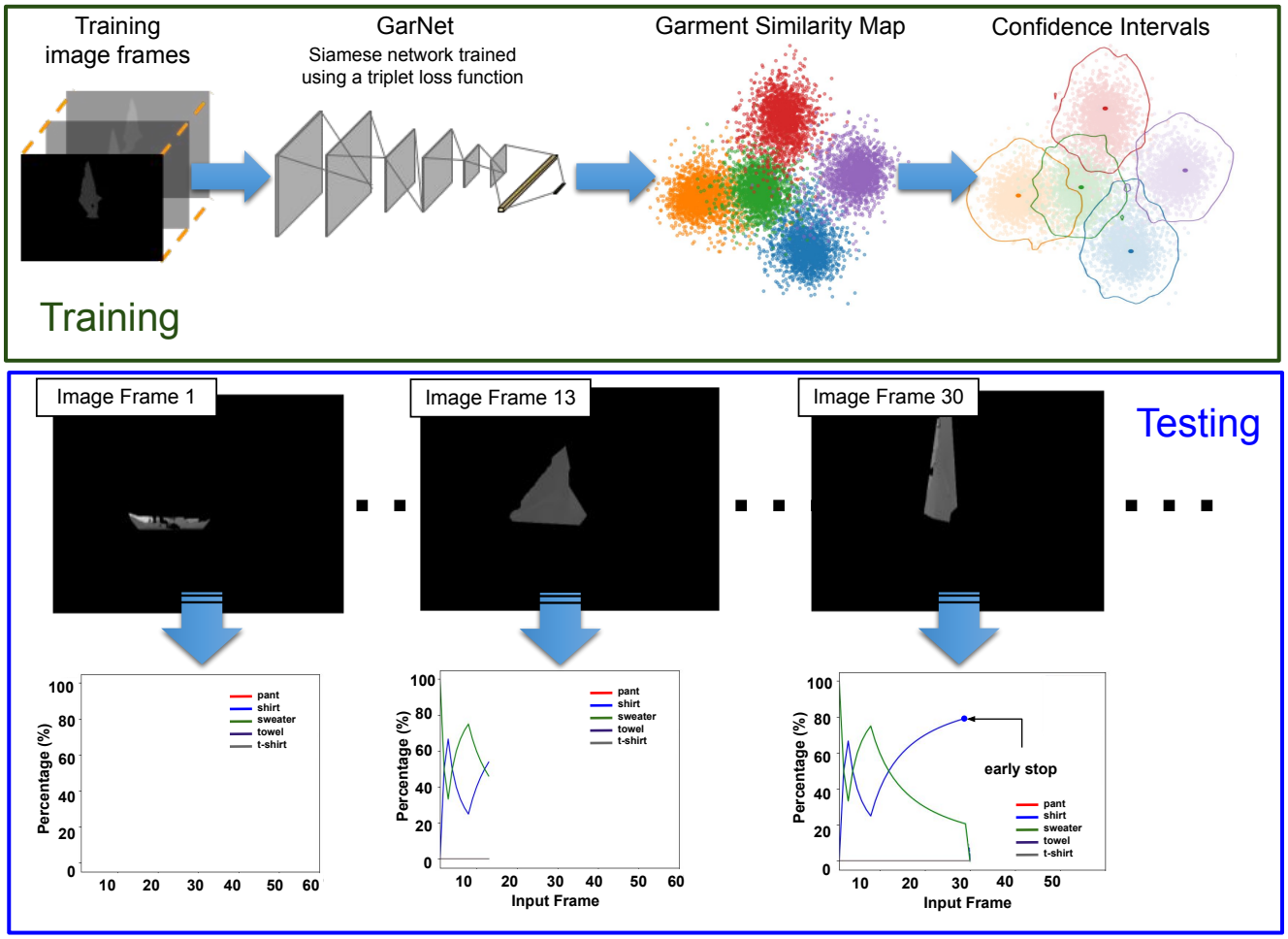

Fig. 2. *Top: Training GarNet.* A positive, negative and anchor image samples from a video sequence of the training dataset are input into GarNet. Training consists of identifying whether any two of the input triplet comes from the same shape or discretised weight categories. GarNet maps input image frames into a Garment Similarity Map (GSM) in which input frames are mapped into clusters if they are similar; otherwise, new points are pulled apart from the cluster. Confidence intervals are computed for each cluster in the GSM as described in Section III-A. *Bottom: Continuous Perception (Testing GarNet).* An image from a video sequence of the testing dataset is input into a trained GarNet to get the mapping onto the garment similarity map. A video sequence of a garment in our database contains 60 frames. The plots show that GarNet gains confidence in predicting that the perceived garment is a shirt. That is, as knowledge is being accumulated into the GSM, most of the decision points belong to the shirt category. In the example shown, the prediction is stopped at frame 30 because 80% of all decision points belong to the shirt category.

1) We have advanced the state-of-art by adopting a continuous perception paradigm in a neural network which improves the prediction accuracy from 70.8% to 93% for shape classification;
2) Our approach can visually estimate weights of unseen garments with a 98.5% prediction accuracy;
3) We propose an early stop strategy so that our approach is faster during inference compared to the state-of-art by taking 8 seconds in average to classify shapes.

II. Literature Review

Previous approaches have proposed learning geometric and physical properties of deformable objects before manipulation. Geometric properties of garments include shapes [4] [5] [3] and physical properties such as weight [6], stiffness (e.g. bending, stretching and shearing) [1], damping factors and elasticity [7].

A. Geometric properties

Maitin-Shepard *et al.* [8] proposed a multi-view grasp-point detection approach for finding grasping points to manipulate towels. Their work demonstrated the importance and effectiveness of learning geometric features of deformable objects before manipulation. However, their approach only enables a robot to manipulate towels rather than garments of various shapes. Similarly, Seita *et al.* [9] proposed a robotic making-bed approach by finding grasping points using a deep neural network. The network was trained on depth images capturing crumpled bed sheets with manually annotated ground-truth grasping points and demonstrated effectiveness in flattening bed sheets in a real environment. However, their method did not include experiments involving multi-shaped garment flattening. Qian *et al.* [10] developed a cloth region segmentation approach to find grasping points to manipulate towels. Their pipeline is divided into three steps: cloth region segmentation, grasp configuration and grasp execution. The key in their work

was finding grasping points using edge and corner detection approaches. They demonstrated their approach by using a robot to find grasping points and grasp the found grasping points. Similar to [8] [9], only garments of one shape are tested.

B. Physical properties

Simulated environments [6] [2] that model deformable objects, have been used in the literature to learn the physical properties of these objects and extrapolate the learned knowledge into the real world. For example, Ruina *et al.* [6] devised an approach where they predicted the area weights of fabrics by learning the physical similarities between simulated fabrics and real fabrics using a spectral decomposition network (SDN). Closing the gap between a simulation and the real world is effective because the physical property parameters of simulated objects are easily accessible compared to those of real objects. However, their approach is not applicable for online evaluation because it requires aligning the simulation with reality, creating an extra overhead before a prediction can be carried out. Hoque *et al.* [2] proposed learning dynamic physical properties of towels via a Vision-Spatial Foresight network (VSF), which is trained on simulated towels but tested on real towels. VSF predicts a sequence of towel deformations and corresponding robot actions based on the towels' initial and desired configurations. They used RGB plus depth images instead of RGB images in their experiment, and they obtained a success rate of 90% on manipulating (flattening) the towels. Even though the robot used to deploy VSF takes unnecessary actions while folding a towel, their proposed approach demonstrates that prior knowledge on understanding geometric and physical properties of deformable objects enables an effective manipulation of deformable objects. Therefore, this paper investigates whether a continuous perception approach coupled with an early-stop strategy can extend beyond simple fabrics and one single shape class. That is, we propose learning the similarity between garments to predict unseen garment shapes and discretised weights based on a 'garment similarity map'. Compared to previous works (e.g. [5], and [3]), our work features an early-stop strategy, where a prediction can be halted earlier without observing the entire interaction.

C. Garment shape prediction: from single-shot perception to continuous perception

Mariolis *et al.* [11] proposed to use a CNN network to classify garment shapes, which are rotated by a dual-arm robot. The CNN network learns garment dynamics via episodic depth images and achieves an accuracy rate of 89% after training. However, they train and test the network on synthetic datasets of simulated garments and their validation image set contains the garments that are already in the training dataset, failing to generalise for unseen garments of similar classes. Chi *et al.* [12] proposed estimating poses of garments by completing their shape from single images. Similar to this paper, the authors allowed a robot to grasp and drop garments to learn their poses. However, they only captured images after the garment was grasped and hanging from the robot's gripper. In this paper, we explore a continuous perception paradigm where a network learns to accumulate knowledge by observing video sequences in order to classify garments shapes and weights.

Sun *et al.* [4] presented an approach where local and global features are extracted from single images and are used to predict unseen garment shapes. This approach makes use of local and global visual characteristics of garments, such as wrinkle features, for shape prediction. Compared to [11], their approach does not require interactions with garments, allowing it to be faster to predict shapes and is robust while being presented with unseen samples. However, prediction accuracies are constrained by the inability of the robot to interact with the garments, and no new knowledge can be captured. For this, Sun *et al.* [3] proposed a Gaussian process regression classifier to predict unseen garment shapes while the robot interacted with them. That is, the robot in their experiment shakes or flips and then drops garments on a table to obtain a new state to increase the classification score. If the classification score of a garment is above a threshold, the garment is sorted based on their shape. This approach, therefore, demonstrated that interacting with garments enables an autonomous system to improve its prediction confidence over interactions and leads to higher classification accuracies.

However, [3] captured the garment's state while being static and on a table which results in making the system slow at predicting shapes as it requires multiple interactions. Martinez *et al.* [5] removed this limitation by introducing the concept of continuous perception to enable a robot to predict shapes by continuously observing video frames from an Xtion depth-sensing camera rather than single image frames. They showed higher accuracy in predicting unseen garment shapes compared to [3] and [4]. However, the limitation in [5] is that they let the robot to observe the entire video sequence before a decision can be made, which means that the robot takes a significant amount of time to predict a garment shape category, and this is given by the length of the video. In their work, they sample a garment for approximately 6 seconds which consists of sampling the garment from a crumpled to a hanging state. In this paper, we, therefore, explore the possibility of adopting the continuous perception paradigm to allow a robot to change its manipulation strategy on the fly, i.e as soon as the class of the garment is predicted even if the garment is still being manipulated from a crumpled to a hanging state.

III. GarNet: Garment Similarity Network

Our proposed Garment similarity Network (GarNet) consists of a Siamese network [6], [13] which clusters garments into groups according to their shape and discretised weight categories. In previous work, Siamese networks provided the ability to cluster similar features such as colour features [6] for predicting flag area weights and physics parameters [13]. Thus, the objective of clustering garments in this paper is to learn common geometric and physical features of garments of the same categories. Our GarNet network comprises a residual convolutional block that extracts features from input data and a fully connected layer that maps features onto a 2D *Garment Similarity Map* (GSM). A garment similarity

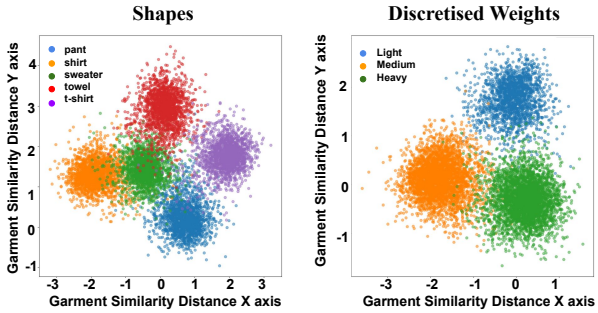

Fig. 3. The Garment Similarity Map (GSM) after training GarNet.

map is a 2D manifold that encodes a garment's geometric and physical characteristics according to its shape and discretised weight categories. Garments of the same categories are clustered together, while garments of different categories are pulled apart. Figure 3 shows the garment similarity map in our experiments where each cluster in the map is called a Garment Cluster ($\mathcal{GC}$). Our GarNet training process is expressed mathematically as $\mathcal{P} = f_\theta(I)$ where f_θ denotes a neural network that contains residual convolutional layers and fully connected layers parameterised by the parameters θ, and I denotes an input video frame. We define $\mathcal{P}$ as a garment similarity point ($\mathcal{GSP}$). Each frame in the input video sequence of the garments is converted into one garment similarity point ($\mathcal{GSP}$). We also define a Garment Similarity Distance ($\mathcal{GSD}$) as $\mathcal{GSD}(x, y) = \mathcal{P}_i - \mathcal{P}_j$; where i and j are the ith and jth garment similarity point. $\mathcal{GSD}$ increases between garments with different labels and decreases between garments with the same labels. Therefore, to train GarNet, we use a triplet loss [14]:

$$\begin{aligned} PP &= \left|\mathcal{P}_{positive} - \mathcal{P}_{anchor}\right| \\ NP &= \left|\mathcal{P}_{negative} - \mathcal{P}_{anchor}\right| \\ TripletLoss &= max(0, PP - NP + margin) \end{aligned} \tag{1}$$

where PP is a positive pair between positive and anchor samples and NP is a negative pair between negative and anchor samples. An anchor sample is an image of a garment. A positive sample is an image of a garment of the same category as the anchor sample. A negative sample is an image of a garment of a different category to the anchor sample. The $margin$ is a value that promotes the network to learn to map positive and negative samples further away from each other. We set this $margin$ to 1 as suggested in [6].

Two GarNets are trained separately and predict the shapes and discretised weights independently. That is, the GarNet for shape predictions is trained on shape categories, while GarNet for discretised weight predictions is trained on discretised weight categories, e.g. light, medium and heavy weights.

A. Garment Cluster Confidence Intervals

To decide which category (either shapes or discretised weights) a mapped garment similarity point in the similarity map belongs to, we propose to fit statistical confidence intervals to each garment cluster in this map. That is, we define a confidence interval using a non-parametric probability density function for each garment cluster, $\mathcal{GC}$ via a kernel density estimator [15] that is defined as:

$$\begin{aligned} \hat{f}_h(\mathcal{GC}) &= \frac{1}{n}\sum_{i=1}^{n} K_h(\mathcal{P} - \mathcal{P}_i) \\ &= \frac{1}{nk}\sum_{i=1}^{n} K(\frac{\mathcal{P} - \mathcal{P}_i}{h}) \end{aligned} \tag{2}$$

where $\mathcal{GC}$ is the garment cluster, K is a Gaussian kernel, $h > 0$ is a smoothing parameter called bandwidth which regulates the amplitude of confidence intervals, and $\hat{f}_h$ is an estimated probability density function for a garment cluster. We have conducted an ablation study on the confidence interval's bandwidths (h) and results are presented in section V-B. After training a GarNet, the centroid of each garment cluster is defined as:

$$\mathcal{GC}_{mean} = (\frac{1}{m}\sum_{i=1}^{m} x_{\mathcal{P}_i}, \frac{1}{m}\sum_{i=1}^{m} y_{\mathcal{P}_i}) \tag{3}$$

where $\mathcal{GC}_{mean}$ is the mean value of garment similarity points mapped from one garment cluster (in Figure 3) and m is the number of garment similarity points in the cluster.

In our experiments, we directly input unseen image frames of garments acquired by the robot to GarNet to map them into the garment similarity map. To decide the shapes and discretised weights, we define a Decision Point ($\mathcal{DP}$) that is the mean value of garment similarity points ($\mathcal{GSP}$s):

$$\mathcal{DP} = (\frac{1}{n}\sum_{i=1}^{n} x_{\mathcal{P}_i}, \frac{1}{n}\sum_{i=1}^{n} y_{\mathcal{P}_i}) \tag{4}$$

where n is the total number of frames observed. To predict the shapes and discretised weights, we find whether a $\mathcal{DP}$ is within any confidence interval and has the minimum distance to the confidence interval's $\mathcal{GC}_{mean}$. For this paper, we use the Euclidean distance to evaluate how close a $\mathcal{DP}$ is with respect to $\mathcal{GC}_{mean}$.

Each video sequence has 60 frames (6 seconds); therefore, we will have 60 decision points. To predict the shape and discretised weight, we establish that a predicted category should have at least 80% of decision points belonging to a garment cluster. If none of the categories fulfils this requirement, we denote that the observed garment does not have a known class. That is, if a decision point is outside any confidence interval, the network is not confident about which category the input garment belongs to. By clustering garments and defining confidence intervals, it is possible to define an early-stop strategy to allow a robotic system to stop its execution if it is confident about the garment shape or weight. After observing a number of image frames of a garment (20 images), if any of the trained categories takes 80% of the decision points, the process is terminated, and the category is chosen as the predicted category.

Fig. 4. *The garment database used in our experiments*: Five categories (pants (jeans), shirts, sweaters, towels and t-shirts), and each category has four garment instances. For each garment instance, we show an RGB image frame and its corresponding segmented depth image

IV. Experiments

A. GarNet Architecture

Our GarNet comprises a ResNet18 [16] as a feature extraction and fully connected networks (FC). The FC networks comprise three linear layers, where a PReLU activation layer is placed between adjacent linear layers. The source code for GarNet and experimental scripts are available at https://liduanatglasgow.github.io/GarNet/.

We use an Intel-i7 equipped with an Nvidia GTX 1080 Ti to train the network. We use the Adam optimizer with an initial learning rate of 1×10^{-3}, controlled by a learning scheduler with a decay rate of 1×10^{-1} and a step size of 8 epochs. The network is trained for 270K iterations with a batch size of 28, taking approximately 30 minutes.

B. Data Collection and Experiments

The video database in this experiment consists of 20 garments of five different shapes, namely, pants, shirts, sweaters, towels and t-shirts. Figure 4 shows garment samples for each garment instance in our database, i.e. five categories and four garments for each category. For each shape, there are four garments of different colours and materials. We used an electric scale to weigh every garment and divide their weights into three discretised levels, namely, light, medium and heavy weights. Therefore, in these experiments, we do not predict the real weight values of tested garments but predict the discretised weights levels in order to enable a robot sort garments as we do before putting garments into a washing machine. We train one GarNet for shape and one for weights. This is because each encodes knowledge based on the observed features, which result in different, uncorrelated 2D manifolds as can be observed in Fig. 3. For example, pants and sweaters (heavy weights) are close together in Fig. 3(left) but are encoded into heavy in Fig. 3(right) which does not correlate to the shape manifold if we merge their clusters.

To validate our network and test our hypothesis (Section I), we propose to carry out a leave-one-out cross-validation methodology. That is, we group all garments into four groups and each garment category as shown in Fig. 4, has four different garment instances. Hence, four experiments are conducted, where three groups served as training groups and one group served as a testing group. The testing group only contains image frames of unseen garments, which means these images are not included in the three training groups. We ensure that the garments in the testing group are entirely 'new' and 'unseen' to the robot. We averaged accuracies for each category output from the four experiments and used the testing group to validate the classification performance of our approach. For each of the four experiments, the training group represents 80% of our video sequence database, while the testing group represents 20% of our database.

We have used a Baxter robot to manipulate garments. The Baxter robot grasped garments from a fixed point, lifted the garments to a point above the table (height is 1m) and then dropped the garments to fall on the table. The running time is 6 seconds where the robot grasps a garment, and stops in the air for 2 seconds before dropping the garments off to the table. An Xtion depth-sensing camera is used to capture garment video sequences. Each garment is captured ten times, which means that the grasp-and-drop operation is conducted ten times. There are 200 videos in total, and each video contains 60 frames (sampling frequency is 10Hz; video sequence length is 6 seconds). Therefore there are 12,000 image frames in total. Figure 1 shows the experimental setup of the robot grasping and dropping garments.

Our experiments include 50 unseen garment videos containing ten videos for each of the four leave-one-out cross-validation experiments. For each video sequence, we predict the shape and discretised weight of the garment in the video. Therefore, we have ten predictions for each category (one prediction for each video) and 50 predictions in total. The prediction accuracy for each category is defined as the percentage of correctly predicted videos of each category.

We compare our approach with four state-of-art (SOA). Duan *et al.* [17] proposed classifying shapes and discretised weights by leveraging a convolutional neural network and a long-term short memory unit (CNN-LSTM). Sun *et al.* [3] provided an interactive approach to classifying garments based on a multi-class Gaussian-Process classifier where a robot gains confidence in predicting garment shapes by shaking and flipping the garments. In their later project [4], they propose to classify garment shapes with a global-local-features classifier, where the classifier captures two local features: local B-Spline Patch (BSP) and locality-constrained linear coding, and three global features: Histogram of Shape Index (SI), Histogram of Topology Spatial Distances (TSD), and Histogram of Local Binary Pattern (LBP). Martinez *et al.* [5] introduced a continuous perception method to classifying the shapes of garments, where a robot observes video sequences being grasped and

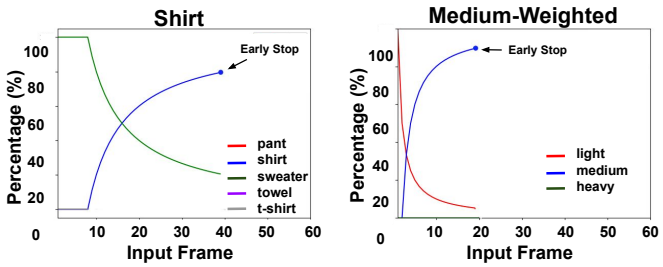

Fig. 5. Examples of the early-stop strategy proposed in this paper. As observed in both plots, GarNet becomes confident over time, and the early-stop strategy activates if 80% of decision points in the garment similarity map are within a correct category. Plots for all unseen predictions can be accessed at https://liduanatglasgow.github.io/GarNet/

dropped and makes decisions on garment shapes based on these video sequences. We compare our approach with these four approaches on shape classification accuracy and running time if reported.

V. Results

While training GarNet, our approach achieves a validation classification performance of 93.9% for shapes and 94.9% for discretised weights. Figure 3 shows the mappings of testing garments onto the similarity map, where it is possible to observe that garments of different categories are pulled apart; while garments of the same categories are clustered together. These results confirm that GarNet coupled with a triplet loss function (Eq. 1) is able to extract physical dissimilarities between categories while maintaining inter-class physical properties within well-defined clusters.

A. Continuous Perception Experiments

Examples of our experimental results are shown in Figure 5. We can see that although GarNet does not recognise garments correctly from the beginning (because most of the decision points belong to an incorrect category), GarNet gradually gains confidence in predicting a correct category for each garment because more decision points are within the correct category and percentages of decision points are eventually over 80%. The classification task is consequently stopped early, and the system does not need to observe the full video sequence to make a correct prediction of the garment class. We conduct two ablation studies for the continuous perception experiment. The first study is about comparing predictions only on local garment similarity points ($\mathcal{GSP}$s) rather than on decision points ($\mathcal{DP}$s). The second ablation study compares the performance of GarNet trained on RGB and depth images. Tables I and II show the results of the leave-one-out cross-validation experiments, where the network achieved 93% for shape classification and 96% for discretised weight classification. The results show that our network has an expected ability to classify shapes and discretised weights of unseen garments.

We use decision points, $\mathcal{DP}$s, (Eq. 4) to make predictions on unseen garment shapes and discretised weights. That is, the position of a decision point on the garment similarity map (in Figure 3) depends on all previously observed image frames rather than on currently observed image frames. From Tables I and II, we can observe that using decision points has better performance than using garment similarity points (93% vs 78% for shapes and 98.5% vs. 80% for discretised weights, respectively). This shows that GarNet benefits from using accumulated knowledge via decision points rather than episodic knowledge as in [4], [11].

To investigate whether the type of image affects the overall prediction of a garment class, we trained GarNet using RGB and depth images. Tables I and II show that a GarNet trained on depth images outperforms a GarNet trained on RGB images and a GarNet trained on RGBD images (93% vs 53.5% vs 47.5% for shapes and 98.5% vs 65% vs 58.5% for discretised weights, respectively). The increase in performance is because depth images capture structural and dynamic information of the garment being manipulated and are better suited to capture the physical properties of garments as opposed to RGB images as proposed by [6]. We have explored this phenomenon in our previous work [13] and found that RGB images capture visual texture information of garments, but this information is affected by lighting conditions that vary between experiments, resulting in worse performance than depth images. Furthermore, texture information from the RGB images is not constant because accessories of garments, colours and lighting conditions quickly change across different garments. We can observe that the GarNet trained on the RGBD dataset unperformed compared to only training on depth or RGB images in Tables I and II. We can conclude that RGB and depth images capture different features of garments and the combination of them makes GarNet to loose the ability (as suggested by [18]) to learn distinctive features that can be used for the shape and weights classification task.

Note that the intra-class variability for the pants category is consistent (i.e. we use jeans for this category, see Fig. 4, top row). Therefore, classification scores in Table I for pants are high across the ablation studies with respect to other shape categories of which they have high intra-class variability. This result shows that in order to generalise to unseen garments, depth images and decision points offer the best combination for the continuous perception task.

B. Ablation study on the confidence intervals bandwidths

A bandwidth, as defined in section III-A, determines the size of a confidence interval. We, therefore, evaluate the effect of the bandwidth selection with respect to the performance of GarNet. A confidence interval of a garment cluster is a region in the garment similarity map of which a certain percentage of $\mathcal{GSP}$s are grouped together.

A decrease in the bandwidth value denotes a decrease in the percentage of $\mathcal{GSP}$s included within the garment cluster. An increase in the bandwidth means that almost all $\mathcal{GSP}$s should be included, while a small portion of points are relatively far away from a cluster. This means that a confidence interval may overlap with other confidence intervals, or even multiple confidence intervals will be generated for one garment cluster. The the final classification prediction depends on the bandwidth.

TABLE I
TABLE: PREDICTION RESULTS (SHAPES)

Category	depth, $\mathcal{DP}$	depth, $\mathcal{GSP}$	RGB, $\mathcal{DP}$	RGB, $\mathcal{GSP}$	RGBD, $\mathcal{DP}$	RGBD, $\mathcal{GSP}$
pants	97.5%	82.5%	97.5%	87.5%	57.5%	10.0%
shirts	77.5%	75%	87.5%	62.5%	55.0%	80.0%
sweaters	97.5%	85%	25%	20.0%	60.0%	5.0%
towels	92.5%	65%	50%	17.5%	17.5%	0.0%
t-shirts	100%	82.5%	32.5%	22.5%	47.5%	37.5%
Average	**93.0%**	78.0%	53.5%	42%	47.5%	26.5%
Standard Deviation	8.1%	**7.3%**	29.5%	28.1%	15.6%	29.7%

TABLE II
PREDICTION RESULTS (DISCRETISED WEIGHTS)

Category	depth, $\mathcal{DP}$	depth, $\mathcal{GSP}$	RGB, $\mathcal{DP}$	RGB, $\mathcal{GSP}$	RGBD, $\mathcal{DP}$	RGBD, $\mathcal{GSP}$
lights	97.5%	50%	37.5%	5.0%	90.0%	55.0%
mediums	97.5%	87.5%	72.5%	58.75%	48.75%	65.0%
heavies	100%	87.5%	71.25%	71.25%	52.5%	18.75%
Average	**98.5%**	80.0%	65.0%	53.0%	58.5%	44.5%
Standard Deviation	**1.2%**	15.0%	13.8%	24.6%	15.8%	21.3%

TABLE III
BANDWIDTH ABLATION STUDY.

Bandwidth	Shapes	Discretised weights
10%	2.0%	4.0%
25%	16.0%	26.5 %
50%	46.0%	26.5%
75%	84.5%	79.5%
95%	**93.0%**	98.5%
99%	82.0%	**99.0%**

TABLE IV
COMPARISONS WITH THE STATE-OF-ART. THE RUNNING TIME IS GIVEN IN SECONDS AND *NA* MEANS NOT AVAILABLE

Method	Accuracy (%)	Time
CNN-LSTM (classification) [17]	48%	17
Interactive Perception [3]	64.2%	*NA*
Single-shot category recognition [4]	67.0%	180
Continuous Perception in [5]	70.8%	6
GarNet (Continuous Perception, Ours)	**93.0%**	**8**

In Table III, we find that a bandwidth of 95% has the best performance, and we use 95% as the bandwidth for the rest of the experiments. However, at a 99% bandwidth, we can observe that the prediction accuracy drops while classifying shapes; this is because $\mathcal{GSP}$s are grouped into incorrect categories.

C. Comparison with the state-of-art methods

We have compared our results with the results from [4] [5] [3] [17]. The garments use in this paper are the same as those reported in [3] [5] [4] [17] in order to ensure fair comparison between approaches. As observed in Table IV, our GarNet outperforms previous work on predicting unseen garment shapes. Also, compared with [17] (17 seconds) and [4] (180 seconds), our method is also faster (8 seconds) because the robot continuously perceives garments without interruptions. There are several reasons why our network has the best performance and we discuss these below.

1) The use of a garment similarity map to encode knowledge of garment shapes and weights: Inspired by [6], where the authors proposed learning the physical similarity between simulated fabrics and predicting physical properties of real fabrics, we find that the similarity network effectively predicts unseen deformable objects, such as garments and fabrics. In our previous work [17], where we focused on utilising solely classification approach rather than the clustering approach in our network, we found that our clustering approach presented in this paper improves over the classification approach. Compared with a traditional classifier that regresses embeddings of data into labels (which is equivalent to asking which shape/discretised weight classes the data belongs to), our GarNet network learns geometric and physical characteristics that makes them the same or different (which is equivalent to asking why the data presents the same or different shapes/discretised weights). Therefore, for unseen garments, the network only needs to decide similarities of the garments for each garment cluster rather than classify them into certain classes, reducing the prediction difficulty.

2) Continuous Perception.: Traditional methods such as [4], [11] that predict shapes and weights of garments are based on static garment features such as wrinkles, outlines, creases, to name a few. Instead, we propose to carry out predictions on encoded knowledge in the GSM while learning the dynamic properties of garments.

3) Early-Stop strategy.: Compared with [3] [5], where the proposed approaches consist of having a robot observing the entire interaction with a garment, our approach only requires a robot to observe interactions partly if termination requirements are satisfied. Therefore, our approach has a mechanism to stop a manipulation on the fly as GarNet can process images every 0.1 seconds, taking an average of 4 seconds to generate a prediction.

VI. CONCLUSION

We have presented a garment similarity network (GarNet) that learns the similarity of the garments and predicts continuously their shapes and their visually perceived weights. We have also introduced a Garment Similarity Map ($\mathcal{GSM}$) that

encodes garment shapes and weights knowledge into clusters. These clusters were then used to decide on which cluster unseen garment samples belongs to heuristically. Our experimental validation shows that, GarNet obtains high prediction accuracies while classifying shapes (93%) and discretised weights (98.5%), Fig. 5. Similarly, we have compared GarNet's performance with the state of the art, and GarNet showed an increase of 22.2% of classification accuracy performance (Table IV).

Compared with previous work on continuous perception [5], GarNet has the advantage of an 'early stop' strategy. That is, a robot does not need to observe the full motion (video sequences) to make predictions which could enable robots to be more responsive and effective while manipulating garments and deformable objects in a laundry pipeline. However, GarNet, in this paper, does not support online learning of unknown garment shapes. For instance, we train GarNet on five shape categories, and it can predict shapes of unseen garments from those categories. Enabling a robot to recognise garments of unknown categories is pivotal in our future work. Currently, our approach only supports classifying garments into known categories. To extend Garnet to unknown categories, we propose to implement a novelty detection approach and using the GSM to detect whether the observed garment is unknown based on the distance from the known clusters. Then, a continual learning approach (e.g. [19]) can be adopted to allow GarNet to be retrained without loosing previous knowledge.

In future work, we plan to devise an online-learning approach for GarNet to investigate complex manipulations interactions (enabled by a behaviour-based reinforcement learning agent [20]), such as twisting garments, shaking garments or rotating garments. From those interactions, differences in stretching and bending characteristics of garments can be exploited to evaluate garments' stiffness parameters, which can potentially help to develop a robot dexterous garment manipulation approach for folding [21], flattening [22], to name a few, that requires fewer iterations. Furthermore, knowledge of the garments weights, i.e. whether it is light, medium or heavy, can enable a robot to plan for these complex manipulation interactions since it will be possible to reduce the search space while estimating the dynamic physical properties of garments.

VII. Acknowledgement

We would like to thank George Killick and Nikos Pitsillos for their valuable comments while reviewing this paper.

References

[1] H. Wang, J. F. O'Brien, and R. Ramamoorthi, "Data-driven elastic models for cloth: modeling and measurement," *ACM transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–12, 2011.

[2] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, "Visuospatial foresight for multi-step, multi-task fabric manipulation," 2021.

[3] L. Sun, S. Rogers, G. Aragon-Camarasa, and J. P. Siebert, "Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 2464–2470.

[4] L. Sun, G. Aragon-Camarasa, S. Rogers, R. Stolkin, and J. P. Siebert, "Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6699–6706, 2017.

[5] L. Martínez, J. R. del Solar, L. Sun, J. P. Siebert, and G. Aragon-Camarasa, "Continuous perception for deformable objects understanding," *Robotics and Autonomous Systems*, vol. 118, pp. 220 – 230, 2019.

[6] T. F. H. Runia, K. Gavrilyuk, C. G. M. Snoek, and A. W. M. Smeulders, "Cloth in the wind: A case study of physical measurement through simulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 495–10 504.

[7] B. Tawbe and A. Crétu, "Acquisition and neural network prediction of 3d deformable object shape using a kinect and a force-torque sensor †," *Sensors (Basel, Switzerland)*, vol. 17, 2017.

[8] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. 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*, 2010, pp. 2308–2315.

[9] D. Seita, N. Jamali, M. Laskey, A. K. Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. Canny, and K. Goldberg, "Deep transfer learning of pick points on fabric for robot bed-making," in *Robotics Research*, T. Asfour, E. Yoshida, J. Park, H. Christensen, and O. Khatib, Eds. Cham: Springer International Publishing, 2022, pp. 275–290.

[10] J. Qian, T. Weng, L. Zhang, B. Okorn, and D. Held, "Cloth region segmentation for robust grasp selection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9553–9560.

[11] I. Mariolis, G. Peleka, A. Kargakos, and S. Malassiotis, "Pose and category recognition of highly deformable objects using deep learning," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 655–662.

[12] C. Chi and S. Song, "Garmentnets: Category-level pose estimation for garments via canonical space shape completion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3324–3333.

[13] L. Duan, L. Boyd, and G. Aragon-Camarasa, "Learning physics properties of fabrics and garments with a physics similarity neural network," 2021. [Online]. Available: https://arxiv.org/abs/2112.10727

[14] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*. Springer, 2015, pp. 84–92.

[15] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832 – 837, 1956.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] L. Duan. and G. Aragon-Camarasa., "Continuous perception for classifying shapes and weights of garments for robotic vision applications," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, 2022, pp. 348–355.

[18] M. B. Shaikh and D. Chai, "Rgb-d data-based action recognition: a review," *Sensors*, vol. 21, no. 12, p. 4246, 2021.

[19] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[20] A. Pore and G. Aragon-Camarasa, "On simple reactive neural networks for behaviour-based reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7477–7483.

[21] A. Doumanoglou, J. Stria, G. Peleka, I. Mariolis, V. Petrik, A. Kargakos, L. Wagner, V. Hlaváč, T.-K. Kim, and S. Malassiotis, "Folding clothes autonomously: A complete pipeline," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1461–1478, 2016.

[22] L. Sun, G. Aragon-Camarasa, S. Rogers, and J. P. Siebert, "Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 185–192.