Cross-domain representation learning for clothes unfolding in robot-assisted dressing

Qie, Jinge; Gao, Yixing; Feng, Runyang; Wang, Xin; Yang, Jielong; Dasgupta, Esha; Chang, Hyung Jin; Chang, Yi

DOI:
10.1007/978-3-031-25075-0

Cross-Domain Representation Learning for Clothes Unfolding in Robot-assisted Dressing

Jinge Qie[1], Yixing Gao[2*], Runyang Feng[2], Xin Wang[2], Jielong Yang[2], Esha Dasgupta[3], Hyung Jin Chang[3], and Yi Chang[2*]

[1] College of Computer Science and Technology, Jilin University, Changchun, Jilin Province, China
qiejg19@mails.jlu.edu.cn
[2] School of Artificial Intelligence, Jilin University, Changchun, Jilin Province, China
gaoyixing@jlu.edu.cn, runyang2019.feng@gmail.com, wxin21@mails.jlu.edu.cn, jyang022@e.ntu.edu.sg, yichang@jlu.edu.cn
[3] School of Computer Science, University of Birmingham, Birmingham, UK
EXD949@student.bham.ac.uk, H.J.Chang@bham.ac.uk

Abstract. Assistive robots can significantly reduce the burden of daily activities by providing services such as unfolding clothes and assistive dressing. For robotic clothes manipulation tasks, grasping point recognition is one of the core steps, which is usually achieved by supervised deep learning methods using large amounts of labeled training data. Given that collecting real labeled data is extremely labor-intensive and time-consuming in this filed, synthetic data generated by physics engines is typically adopted for data enrichment. However, there exists an inherent discrepancy between real and synthetic domains. Therefore, effectively leveraging synthetic data together with real data to jointly train models for grasping point recognition is desirable. In this paper, we propose a Cross-Domain Representation Learning (CDRL) framework that adaptively extracts domain-specific features from synthetic and real domain respectively, before further fusing these domain-specific features to produce more informative and robust cross-domain representations, thereby improving the prediction accuracy of the grasping points an assistive robot must take advantage of. Experimental results show that our CDRL framework is capable of recognizing grasping points more precisely than when compared with five baseline methods. Based on our CDRL framework, we enable a Baxter humanoid robot to unfold a hanging white coat with a 92% success rate and to successfully assist 6 users in dressing.

Keywords: Clothes unfolding · Grasping point recognition · Robot-assisted dressing · Human-robot interaction

1 Introduction

In our daily life, dressing is an important activity in which many people need assistance due to disabilities or impairments. [17]. Assistive robots can help with

*Corresponding author

Fig. 1: The Baxter humanoid robot automatically recognizes the hanging clothes' grasping points and then unfolds the clothes to a wearable state so as to assist users with dressing. The grasping points of the clothes are recognized by our proposed Cross-Domain Representation Learning framework.

reducing the burden of dressing. In recent years, interest has increased in the challenge of robot-assisted dressing [22,10,11,23,9,27,12], they attempt to alleviate the dressing burden via diverse techniques. However, these works mainly focus on the process of dressing and simplify the initial configuration, which usually leads to assumption that the clothes has already been grasped by a robot.

Considering the clothes are often in a hanging state before the dressing starts, a robot should unfold them to wearable states. In robotic clothes unfolding research, the precise recognition of grasping points is fundamental to the performance. Earlier work focused on the use of a random forest algorithm [8] or a clothes template matching method [19] to recognize the clothes' grasping points. With the emergence of deep learning, researchers [5,25,27] utilized Convolutional Neural Networks (CNN) to learn the Cartesian coordinates of grasping points from large-scale labeled data. The performance of deep learning relies heavily on large-scale labeled data, but in the field of robotics, the real labeled data acquisition is extremely time-consuming and labor-intensive. Therefore, employing physics engines to generate synthetic images to augment training datasets has become a widely-adopted paradigm in robotic clothes unfolding tasks [5,26,25,27].

However, due to the inherent discrepancy between the real and synthetic domains, it can be observed that directly applying synthetic images in the training process only improves the model's performance slightly [27].

In this paper, we present a Cross-Domain Representation Learning (CDRL) framework that sufficiently extracts knowledge from both synthetic and real domain to produce more robust cross-domain generalized representations. The CDRL network consists of two main modules. A *Domain-specific Feature Refinement Module* adopts ResNet-101 [14] as a backbone to extract vanilla image features which are domain-irrelevant, then the features are adaptively refined by two domain-aware deformable convolutional [7] branches to produce domain-specific knowledge. A *Cross-Domain Representation Fusion Module* fuses the features of two domain branches to acquire cross-domain representation, this integrates the domain-specific knowledge to improve the model accuracy.

Extensive experiments demonstrate that the proposed CDRL framework significantly outperforms other baseline methods [25,21,5,27] in terms of clothes grasping point recognition (three for single domain methods, two for mixed domains methods). Moreover, we also achieve a 92% robotic clothes unfolding success rate in a real lab environment and enable a Baxter robot to successfully assist 6 real users with dressing.

The main contributions of this paper can be summarized as follows:

- We examine the robotic clothes unfolding task from the perspective of cross-domain representation learning for the first time, aiming to effectively leverage the synthetic data that is easily accessible.
- We propose a Cross-Domain Representation Learning (CDRL) framework for the recognition of clothes' grasping points which can fully extract cross-domain representations through both synthetic and real domain data, and improve the grasping point recognition accuracy.
- Empirical results demonstrate that the proposed CDRL framework can accurately recognize grasping points. We further enable a Baxter robot to bimanually unfold the hanging clothes to a wearable state and assist users with dressing.

2 Related Work

2.1 Robotic Clothes Unfolding

In robotic clothes unfolding tasks, precise grasping point recognition is crucial to the clothes unfolding performance. Earlier studies used manual feature extraction methods for detecting the clothes, such as shapes [6], volumes [20], edges and corners [15] to determine where to grasp. Doumanoglou et al. built random forests based on a clothes depth image dataset that was manually taken and labelled, which was a very expensive and time-consuming approach to implement in practice [8]. Kita et al. proposed a model-driven approach, which used a 3D clothes model to identify the state of the real clothes by matching templates predefined in the generated simulated clothes database [16].

Currently, researchers typically use Convolutional Neural Networks (CNN) for grasping point recognition in robotic clothes unfolding tasks [5,25,27]. However, deep learning models rely on large-scale, high-quality labeled data to exploit efficient feature representation capacity [18]. In many works, especially in the field of robotics, the acquisition of real labeled training data is time-consuming and arduous. Synthetic data generated from physics engines, due to its ease of acquisition and labeling properties, has been used as a means of data augmentation in robotics research, including visual space recognition [26] and navigation [24]. In robotic clothes-related tasks, researchers use synthetic data generated by physics engines and leverage CNN models to learn Cartesian coordinates of a grasping point from large-scale labeled data. Corona et al. [5] and Saxena et al. [25] augment the real dataset with synthetic data and proposed multi-layer convolutional networks to predict the grasping point coordinates. Similarly, Zhang et al. used the AlexNet model to regress single point coordinates from a synthetic and real domain clothes dataset, which enabled the robot to successfully grasp a single clothes point and put one sleeve onto the user arm [27].

These aforementioned works have made advances in clothes grasping point recognition, but the natural domain discrepancy between the synthetic and real domain makes them unable to adequately extract cross-domain generalized representations, thus undermining the model's performance.

2.2 Robot-assisted Dressing

Providing dressing assistance remains an important but challenging problem for robots. Recently, there have been a growing number of research on robot-assisted dressing. Reinforcement learning algorithms [22,3] and demonstration learning methods [2,23] are adopted to teach the robot to learn the dressing motions. In user modeling aspects, user preference has been considered to enable the robot to personalize the dressing assistance for users who suffer from disabilities or impairments [10]. On the other hand, multi-modal information integration allows the robot to perceive users more precisely, thus making the dressing process more efficient and reliable [9]. However, the above research mostly focused on the dressing process, which usually assumed that the clothes had already been grasped by a robot in the configuration setup. In this work, we consider the step of robotic clothes unfolding before the robot-assisted dressing process.

3 Cross-Domain Representation Learning

In this paper, we propose a Cross-Domain Representation Learning (CDRL) framework which adaptively extracts domain-specific features from synthetic and real domain and then fuses the features to yield cross-domain representations. The overall pipeline of the CDRL framework is illustrated in Fig. 2, which consists of a Domain-specific Feature Refinement Module and a Cross-Domain Feature Fusion Module. In the training phase, the CDRL takes a labeled depth image dataset as input which includes both synthetic images and real images.

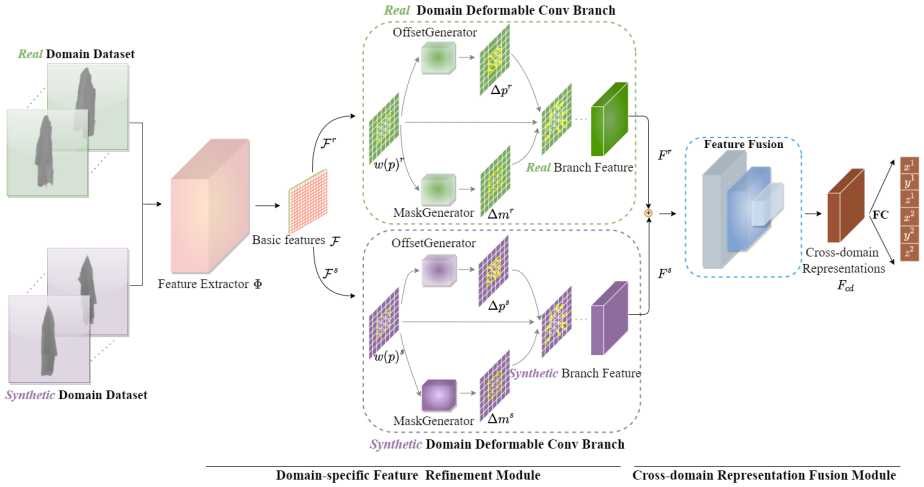

Fig. 2: Overall pipeline of the proposed Cross-Domain Representation Learning (CDRL) framework. This framework takes a depth images dataset (drawing from both the synthetic and real domain) as input in the training phase. The CDRL framework consists of two main modules, a Domain-specific Feature Refinement module which includes a backbone feature extractor Φ to extract basic features $\mathcal{F}$, and then the features $\mathcal{F}^r$, $\mathcal{F}^s$ tagged with their corresponding domains are fed into the domain-aware deformation convolutional branches to adaptively refine and attain domain-specific representations F^r,F^s, where the superscripts $\{r, s\}$ represent the feature from the real and synthetic domain respectively. The Cross-Domain Representation Fusion Module integrates the two domain-specific features and attains cross-domain representations. The Fully-connected (FC) layer transforms the fused representations to the grasping point coordinate outputs. Best viewed in color.

The clothes depth image dataset acquisition and labeling will be described in section 4. We now introduce the two components of CDRL in detail.

Domain-specific Feature Refinement Module Clothes are typically non-rigid objects with complex surface deformations, which are intractable in clothes grasping point recognition. However, the traditional convolution operation adopts a fixed structure that is insufficient for modeling the highly complex nature of deformable clothes. As a result, we leverage the Deformable Convolutional Network (DCN) [28] for adaptively extracting domain-specific representations due to its remarkable transformation modeling capacity.

In particular, given a synthetic or real depth image of clothes, we first employ the pretrained ResNet-101 [14] as the backbone denoted as Φ to extract vanilla features $\mathcal{F}$, which are domain-irrelevant. Then, the features $\mathcal{F}$ are fed into the domain-aware deformable convolution branches, in which the sampling location

weight w(p), offset Δp and modulation scalar Δm are the parameters that need to be learned. These parameters are computed as follows:

$$\begin{aligned} \Delta p^d &= \text{OffsetGenerator}(\mathcal{F}^d), \\ \Delta m^d &= \text{MaskGenerator}(\mathcal{F}^d), \end{aligned} \tag{1}$$

where $d = \{r, s\}$ denotes the real or synthetic domains, the *OffsetGenerator* and *MaskGenerator* are the two separate 3×3 convolutional structures.

Compared to the fixed traditional convolution operation, in a deformable convolution network, the adaptive learnable offset Δp and the modulation scalar Δm are added. With the sampling location grid $\mathcal{K} = \{(1,-1),(0,-1),...,(1,0),(1,1))\}$, the output domain-specific feature map $y(p)$ in the deformable convolutional branches is expressed as:

$$y(p) = \sum_{k=1}^{K} w(p_k) \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{2}$$

where Δp_k and Δm_k denote the offset and modulation scalar at the k-th location in $\mathcal{K}$, respectively. The $x\left(p + p_k + \Delta p_k\right)$ is a bilinear interpolation to prevent sampling offsets from getting fractional values. With the help of these parameters, the deformable convolutional operation can effectively obtain useful location cues from the vanilla features $\mathcal{F}$ and better adapt to the different target domains' features, thus generating high-quality domain-specific features. The above operation can be expressed as:

$$(\mathcal{F}^d, p^d, m^d) \xrightarrow[\text{Convolution}]{\text{Modulated} \quad \text{Deformable}} F^d. \tag{3}$$

Cross-Domain Feature Fusion Module In this module, the domain-specific features F^{r} and F^{s} are concatenated and then passed to several regular 3×3 convolutions for aggregation, which produces the cross-domain representations F_{cd}. Note that F_{cd} integrates the knowledge from both the real and synthetic domains, which is favorable for subsequent grasping point regression. Ultimately, we employ a fully connected layer to decode the final positions of grasping points from F_{cd}.

Loss Function We adopt the mean square error (MSE) to supervise the learning of final grasping points recognition. The loss function is defined as:

$$L\left(\theta\right) = \alpha \cdot \text{MSE}(P_1, T_1) + (1-\alpha) \cdot \text{MSE}(P_2, T_2) + \beta \Omega\left(\theta\right), \tag{4}$$

where P_1, P_2 denote the predicted Cartesian coordinates of the two predicted grasping points. The MSE calculates the error distance between P_1, P_2 and ground truth positions of grasping points T_1, T_2. The α is a hyperparameter used to balance the loss item of each predicted grasping point, and the regularization term $\Omega(\theta)$ is used to alleviate overfitting.

4 Data Acquisition

In order to train an accurate clothes grasping point recognition model, a substantial volume of high-quality labeled training data is necessary. Depth maps are desirable due to their invariance to different colors and textures. In real lab, it is labor-intensive and time-consuming to collect real depth images and label the point coordinates, hence we utilize a physics engine Maya [1] to simulate real lab settings and generate large-scale labeled training data. The acquired real and synthetic depth image samples of the clothes are shown in Fig. 3.

Real Data: As shown in Fig. 1, in our lab setting, we position a rail in front of the robot and hang a white coat randomly on the rail, while a Kinect v2 camera is placed on the left side of the Baxter robot, which is 60 cm down and 100 cm back from the hanging clothes. We gather real depth images by constantly changing the hanging positions of the white coat with the help of the Kinect v2 camera. While taking depth images, the spatial Cartesian coordinates of the grasping points are recorded with a NOKOV Motion Capture System by placing markers at the collar areas. After repeating the above steps, we obtain a total of 5000 pieces of real labeled data, which takes approximately 50 hours. The non-clothing segments are filtered from the real images by thresholding the depth between 80cm and 110cm.

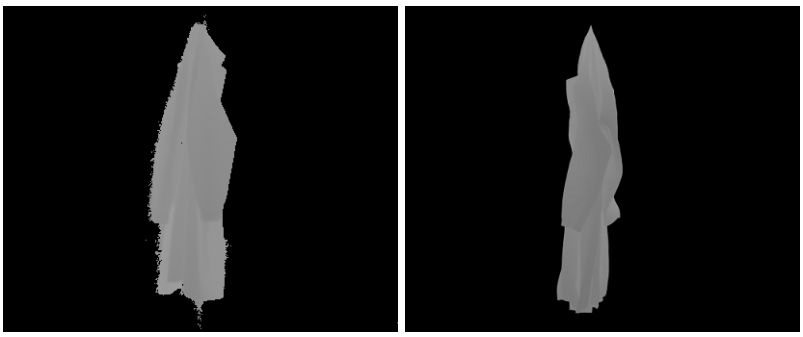

Fig. 3: Samples of real and synthetic depth images. The real depth image (left) is taken by a Kinect v2 camera and the synthetic depth image (right) is generated by the Maya physics engine.

Synthetic Data: We use the physics engine Maya [1] to acquire synthetic clothes images with corresponding grasping points coordinate labels. In Maya, we simulate the real lab environment and set the same relative positions of the camera and the white coat model. In the camera parameter setting, we configure the focal length, horizontal and vertical angle the same as the Kinect v2 camera. Before the data acquisition, we define a number of hanging points on the 3D white coat model to simulate the clothes hanging poses on the rail in the real lab environment. During the acquisition procedure, we simulate the clothes

hanging poses by applying a simulation of gravity at different hanging points. Meanwhile, we alter the attributes of the clothes model, such as compression resistance and bending resistance, to generate diversified data. When the clothes model is stabilized in the gravity simulation, the camera takes a clothes depth image and records the Cartesian coordinates of the predefined grasping points at the collar position. This process is illustrated in Fig. 4. By repeating the above procedure, a total of 14000 labeled depth images are obtained.

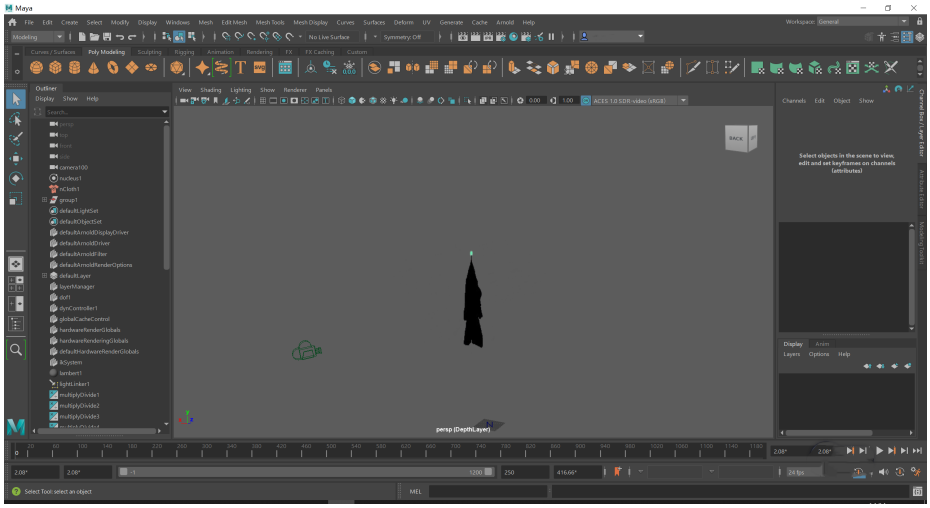

Fig. 4: Maya modeling environment for generating synthetic clothes depth images. In the predefined settings, the clothes object will possess similar features of the real lab white coat, before they are subjected to gravity.

5 Experiments and Results

We first validate the performance of the proposed Cross-Domain Representation Learning (CDRL) framework for clothes grasping point recognition using our collected dataset. Then, based on the proposed CDRL framework, in a real lab environment, we enable the Baxter robot to unfold the hanging clothes and assist users with dressing to further examine the effectiveness of framework.

5.1 Experimental Setup

In lab environment, we set the camera, clothes and Baxter robot to the same as described in Section 4. The Kinect v2 camera captures depth images for grasping point recognition implements human joints tracking algorithm during robot-assisted dressing. The transformation between coordinates has been determined

prior to the experiment. In the CDRL network, we set the learning rate to 0.001, batch size 32. The β and α in Eq. 4 are set to $1e-8$ and 0.4, respectively.

All experiments were conducted on a desktop running Ubuntu 16.04 with a 2.20GHz Intel Xeon Gold 5120 processor and an Nvidia Titan RTX GPU, upon which a ROS operating system and a MoveIt! motion planning library [4] were used to enable the Baxter robot to unfold the coat and assist the user in dressing.

5.2 Approach Evaluation

We conduct extensive experiments with different training dataset settings to evaluate the performance of the CDRL framework for grasping point recognition.

We divide the real data into the training set, validation set and test set with a ratio of 6 : 2 : 2. A total of 14000 synthetic images will be used to collaboratively train the model with an increasing number of real training images $500 \rightarrow 1000 \rightarrow 2000 \rightarrow 3000$. We compare against the following 6 methods:

Single domain methods:

(1) Backbone training with only synthetic data, denoted as *Syn_only*: This baseline corresponds to the approach [25] using only synthetic data.

(2) Backbone training with only real data, denoted as *Real_only* : This baseline aims to train the network using only real images as done in [21].

(3) Backbone training only with noisy synthetic data denoted as *Noised_Syn_only*. Since the depth maps captured by a Kinect v2 camera are noisy, while the synthetic images are very smooth. Therefore, it is desirable to add simulated noise to synthetic images to make them more similar to the real images. Practically, the adopted Kinect noise model [13] uses random offsets to shift pixel locations and adds Gaussian noise, which is corresponded to [5].

Table 1: Single domain methods performance comparisons of (1), (2), (3).

Method	Training data number	Mean Error Distance ↓
Syn_only	14000	5.72 cm
Real_only	3000	1.8 cm
Noised_Syn_only	14000	5.57 cm

Mixed domain methods:

(4) Backbone training on synthetic data with incremental real data $500, 1000, 2000, 3000$, denoted as *Incre_Syn* [27].

(5) Backbone training on noisy synthetic data with incremental real data $500, 1000, 2000, 3000$, denoted as *Incre_Noised_Syn* [27].

(6) Complete CDRL framework training on synthetic data with increasing real images $500, 1000, 2000, 3000$, denoted as *CDRL*.

We verify these methods' prediction accuracy individually using the Mean Error Distance, which measures the error distance between each predicted grasping point and corresponding ground truth coordinates. We provide the results of experimental configurations (1), (2), and (3) in Table 1. For the experimental configurations (4), (5), and (6), the corresponding results are depicted in Fig. 5.

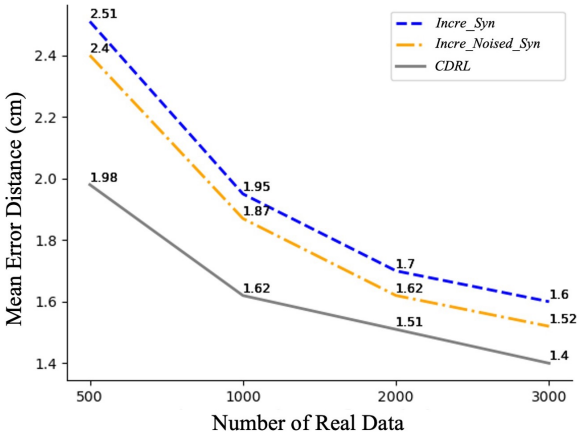

Fig. 5: This figure shows the performance results of mixed domain methods. The Mean Error Distance of the incremental real data learning configurations (4), (5), and (6) on the testset, from which we can see that our CDRL framework outperforms the rest.

From the results depicted in Table 1 and Fig. 5, we can observe that:

(1) The *Syn_only* model (configuration (1)) which trained using only synthetic data has the largest error value (5.72cm). At the same time, the *Real_only* model (configuration (2)) attains 1.8cm error value, which is approximately 31.8% of *Syn_only*. This significant performance gap indicates that there is a clear discrepancy between real and synthetic domains, and directly learning knowledge from the synthetic domain is challenging to transfer to the real domain. On the other hand, for the *Incre_Noised_Syn* (configuration (3)), synthetic training data attached with simulated noise looks more similar to the real images, thereby providing a slightly improvement performance over the configuration (1) that training model using only original synthetic data.

(2) The prediction error of *Incre_Syn* (configuration (4)) gradually decreases to 2.51cm, 1.95cm, 1.7cm, 1.6cm, with the increasing number of real images 500 $\rightarrow$ 1000 $\rightarrow$ 2000 $\rightarrow$ 3000, as illustrated in Fig. 5. Similar trends can be found in (*Incre_Noised_Syn* (configuration (5)) and *CDRL* (configuration (6)), as depicted in Fig. 5. This performance improvement upon the incorporation of real images shows that real images allow the learned distribution close to the real domain, which is favorable for model training.

(3) Remarkably, our proposed *CDRL* achieves state-of-the-art recognition performance and delivers a substantial improvement over all baseline methods, with a final prediction error of 1.4cm. This significant performance demonstrates the effectiveness of cross-domain representation learning. In the *CDRL* framework, through our principled design of the Domain-specific Feature Refinement Module for adaptively extracting domain-specific knowledge and the Cross-Domain Feature Fusion Module for sufficient feature fusion, this framework can obtain robust cross-domain representations and produce the best results.

5.3 Robotic Clothes Unfolding and Assistive Dressing

In this section, based on our proposed CDRL framework, we conduct experiments on a Baxter robot to unfold clothes and assist dressing in a real lab environment.

Robotic Clothes Unfolding Once the clothes grasping points are identified by the proposed CDRL framework, we conduct the robot motion planning using the MoveIt! [4] library to grasp them bimanually from the hanging state to the wearable state. The complete robotic clothes unfolding procedure is shown in Fig. 6. We perform 50 experiments by constantly changing the clothes gestures on the rail, and achieve 92% successful rate of clothes unfolding.

Fig. 6: The entire procedure of unfolding clothes by a Baxter robot. Our CDRL framework calculates the Cartesian coordinate of the clothes grasping points, then the Baxter robot performs motion planning to grasp and unfold the clothes to a wearable state. More detailed video demonstrations can be seen in the supplementary file.

Robot-assisted Dressing With the clothes unfolded by the robot to a wearable state, we remove the rail and users are allowed to stand in front of the robot. The user's arms are held back at a certain angle (30°) to the body as the initial gesture. The Kinect v2 camera SDK based on the camera behind the users will calculate the location of the wrist p_{wst}, elbow p_{elb} and shoulder p_{shd}. Finally, the Baxter robot plans a motion path passing above these key points ($p_{wst} \rightarrow p_{elb} \rightarrow p_{shd}$) to assist users to accomplish the dressing process.

We invited six participants (informed consent was obtained) to get dressed with the help of the Baxter robot. The whole process is illustrated in Fig. 7. The

Fig. 7: Examples of the Baxter robot performing assistive dressing. After the robot unfolds the hanging clothes to a wearable state, the Kinect v2 camera SDK detects the users' joint positions. Following this, the robot performs motion planning to assist users in dressing. More detailed video demonstrations can be seen in the supplementary file.

procedure was performed successfully in most trials, but there exist some failure cases. For example, when participants wear thick or frictional clothes, the robot end-effectors' limited power cannot assist properly and release the clothes.

6 Conclusion

In this paper, we examine the robotic clothes unfolding task from the perspective of cross-domain representation learning for the first time. We present a Cross-Domain Representation Learning (CDRL) framework for clothes grasping point recognition, which adaptively extracts domain-specific features from both synthetic and real domain, and fuses them to produce more robust clothes representations. Experimental results demonstrate that our framework can significantly reduce the mean error of detected grasping points with the same data

volume settings. Furthermore, based on our CDRL framework, we enable the Baxter humanoid robot to unfold the hanging clothes and assist 6 real users in getting dressed. In our future work, we aim to enable our model to support more types of clothes as well as more sophisticated grasping strategies to improve the robot actual performance in real lab experiments.

Acknowledgements The authors would like to thank the anonymous referees for their valuable comments. This work is supported by the National Natural Science Foundation of China (No. 61976102 and No. U19A2065) and the Fundamental Research Funds for the Central Universities, JLU. This work is supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62106082.

References

1. Autodesk, INC.: Maya, `https:/autodesk.com/maya`
2. Canal, G., Alenyà, G., Torras, C.: Personalization framework for adaptive robotic feeding assistance. In: Proceedings of the International Conference on Social Robotics (ICSR). pp. 22–31. Springer (2016)
3. Clegg, A., Erickson, Z., Grady, P., Turk, G., Kemp, C.C., Liu, C.K.: Learning to collaborate from simulation for robot-assisted dressing. Robotics and Automation Letters (RA-L) **5**(2), 2746–2753 (2020)
4. Coleman, D., Sucan, I., Chitta, S., Correll, N.: Reducing the barrier to entry of complex robotic software: a moveit! case study. arXiv preprint arXiv:1404.3785 (2014)
5. Corona, E., Alenya, G., Gabas, A., Torras, C.: Active garment recognition and target grasping point detection using deep learning. Pattern Recognition **74**, 629–641 (2018)
6. Cusumano-Towner, M., Singh, A., Miller, S., O'Brien, J.F., Abbeel, P.: Bringing clothing into desired configurations with limited perception. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 3893–3900 (2011)
7. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 764–773 (2017)
8. Doumanoglou, A., Kargakos, A., Kim, T.K., Malassiotis, S.: Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 987–993 (2014)
9. Erickson, Z., Clever, H.M., Turk, G., Liu, C.K., Kemp, C.C.: Deep haptic model predictive control for robot-assisted dressing. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 4437–4444 (2018)
10. Gao, Y., Chang, H.J., Demiris, Y.: User modelling for personalised dressing assistance by humanoid robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1840–1845 (2015)
11. Gao, Y., Chang, H.J., Demiris, Y.: Iterative path optimisation for personalised dressing assistance using vision and force information. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4398–4403 (2016)

12. Gao, Y., Chang, H.J., Demiris, Y.: User modelling using multimodal information for personalised dressing assistance. IEEE Access **8**, 45700–45714 (2020)
13. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 1524–1531 (2014)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Kampouris, C., Mariolis, I., Peleka, G., Skartados, E., Kargakos, A., Triantafyllou, D., Malassiotis, S.: Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 1656–1663 (2016)
16. Kita, Y., Ueshiba, T., Neo, E.S., Kita, N.: Clothes state recognition using 3d observed data. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 1220–1225 (2009)
17. Lawton, M.P., Brody, E.M.: Assessment of older people: self-maintaining and instrumental activities of daily living. The gerontologist **9**(3_Part_1), 179–186 (1969)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
19. Li, Y., Xu, D., Yue, Y., Wang, Y., Chang, S.F., Grinspun, E., Allen, P.K.: Regrasping and unfolding of garments using predictive thin shell modeling. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 1382–1388 (2015)
20. Li, Y., Yue, Y., Xu, D., Grinspun, E., Allen, P.K.: Folding deformable objects using predictive simulation and trajectory optimization. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6000–6006 (2015)
21. Mariolis, I., Peleka, G., Kargakos, A., Malassiotis, S.: Pose and category recognition of highly deformable objects using deep learning. In: Proceedings of the International Conference on Advanced Robotics (ICAR). pp. 655–662. IEEE (2015)
22. Matsubara, T., Shinohara, D., Kidode, M.: Reinforcement learning of a motor skill for wearing a t-shirt using topology coordinates. Advanced Robotics **27**(7), 513–524 (2013)
23. Pignat, E., Calinon, S.: Learning adaptive dressing assistance from human demonstration. Robotics and Autonomous Systems **93**, 61–75 (2017)
24. Sadeghi, F., Levine, S.: Cad2rl: Real single-image flight without a single real image. arXiv preprint arXiv:1611.04201 (2016)
25. Saxena, K., Shibata, T.: Garment recognition and grasping point detection for clothing assistance task using deep learning. In: Proceedings of the IEEE/SICE International Symposium on System Integration (SII). pp. 632–637 (2019)
26. Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., Essa, I., Parikh, D., Batra, D.: Embodied question answering in photorealistic environments with point cloud perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6659–6668 (2019)
27. Zhang, F., Demiris, Y.: Learning grasping points for garment manipulation in robot-assisted dressing. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 9114–9120 (2020)
28. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9308–9316 (2019)