Grasp-Oriented Fine-grained Cloth Segmentation without Real Supervision

Ruijie Ren*
rren@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain

Mohit Gurnani Rajesh*
mohit.rajesh-gurnani18@imperial.ac.uk
Imperial College London
London, UK

Jordi Sanchez-Riera
jsanchez@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain

Adrian Lopez-Rodriguez
al4415@imperial.ac.uk
Imperial College London
London, UK

Fan Zhang
f.zhang@imperial.ac.uk
Imperial College London
London, UK

Yurun Tian
y.tian@imperial.ac.uk
Imperial College London
London, UK

Guillem Alenyà
galenya@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain

Antonio Agudo
aagudo@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain

Yiannis Demiris
y.demiris@imperial.ac.uk
Imperial College London
London, UK

Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk
Imperial College London
London, UK

Francesc Moreno-Noguer
fmoreno@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain

ABSTRACT

Automatically detecting graspable regions from a single depth image is a key ingredient in cloth manipulation. The large variability of cloth deformations has motivated most of the current approaches to focus on identifying specific grasping points rather than semantic parts, as the appearance and depth variations of local regions are smaller and easier to model than the larger ones. However, tasks like cloth folding or assisted dressing require recognizing larger segments, such as semantic edges that carry more information than points. We thus first tackle the problem of fine-grained region detection in deformed clothes using only a depth image. We implement an approach for T-shirts, and define up to 6 semantic regions of varying extent, including edges on the neckline, sleeve cuffs, and hem, plus top and bottom grasping points. We introduce a U-Net based network to segment and label these parts. Our second contribution is concerned with the level of supervision required to train the proposed network. While most approaches learn to detect grasping points by combining real and synthetic annotations, in this work we propose a multilayered Domain Adaptation strategy that does not use any real annotations. We thoroughly evaluate our approach on real depth images of a T-shirt annotated with fine-grained labels, and show that training our network only with synthetic labels and our proposed DA approach yields results competitive with real data supervision.

CCS CONCEPTS

• **Computing methodologies** → **Vision for robotics**.

KEYWORDS

perception for grasping and manipulation, segmentation and categorization, domain adaptation, deep learning

ACM Reference Format:
Ruijie Ren, Mohit Gurnani Rajesh, Jordi Sanchez-Riera, Adrian Lopez-Rodriguez, Fan Zhang, Yurun Tian, Guillem Alenyà, Antonio Agudo, Yiannis Demiris, Krystian Mikolajczyk, and Francesc Moreno-Noguer. 2023. Grasp-Oriented Fine-grained Cloth Segmentation without Real Supervision. In ,. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Identifying specific regions for grasping is one of the main open challenges in robotic cloth manipulation, due to the large variability of geometric configurations and textures that garments exhibit. Although the effects of texture variability can be reduced by using depth images, which compared to RGB images also present a

*First two authors contributed equally.

,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

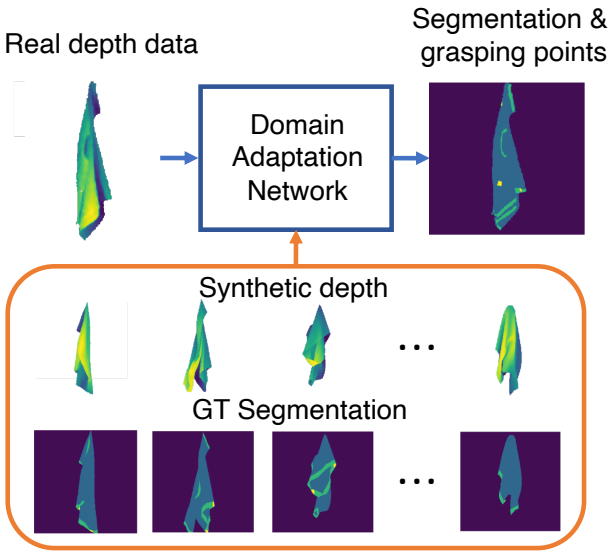

Figure 1: Overview of our approach. We introduce a pipeline for fine-grained semantic segmentation of clothes depth maps. The pipeline leverages a multi-layered Domain Adaptation strategy that allows us to use the proposed network on real depth maps whilst only using synthetic ground truth labels during training. The detected segments are in turn used for improving the performance on the grasping point detection task without using real labels.

lower gap between synthetic and real data [20, 22] that is helpful when leveraging synthetic data for training, the problem is still severely ill-posed. As illustrated in Fig. 1, identifying semantically meaningful regions in the depth map of a garment is challenging even for humans. Most existing methods have therefore focused on identifying key grasping points, as small local regions tend to have more invariant geometric features than large patches [5, 10–14, 17, 21]. Nevertheless, many tasks in cloth manipulation would benefit from detecting larger and more informative regions. In particular, identifying semantic edges (e.g. neckline, sleeve cuffs and hem in a T-shirt) is highly useful for tasks such as folding clothes or robot-assisted dressing.

The first contribution of this work is a method to perform fine-grained edge detection on highly crumpled clothes. We focus on the specific case of a T-shirt, although the approach is generalizable to other garments. We define up to 6 semantic labels of different types and extents, including full body, 3 edge types, and 4 grasping points. We thus adapt a U-Net architecture that given input depth images, can provide the semantic labels of each category (including grasping points, which are treated as regions). The second contribution of our work is addressing the amount of supervision required to train our network. Existing approaches use either a training set of real depth maps annotated with ground truth grasping points or a combination of those with synthetic annotations [5, 10, 11, 17, 21]. In this work, motivated by the difficulty of collecting and annotating real ground truth depth maps with fine-grained edge labels for training, we explore the limits of relying exclusively on supervision from synthetically generated data. For this purpose, we synthesize and annotate several thousands of samples of a T-shirt hanging under gravity from random points. Additionally, we also create a dataset of real depth maps pseudo-annotated with the fine-grained labels to test our methods. We then investigate several training alternatives. More specifically, we propose a Multi-layer Domain Adaptation (DA) approach to reduce the domain gap between the feature maps extracted from synthetic and real data using an adversarial loss computed from non-annotated synthetic/real samples. A thorough evaluation shows that this scheme achieves results competitive with architectures trained on the pseudo-annotated real samples.

In summary, our main contributions are:

- We are the first to tackle the problem of fine-grained edge segmentation in depth maps of highly deformed clothes.
- We explore the limits of Domain Adaptation strategies that leverage uniquely on supervision from synthetic annotations.
- We generate large and realistic synthetic data and collect a mid-size real dataset of deformed T-shirts which we annotated with edge labels and grasping points. This dataset can be used either for finetuning synthetically trained networks or for evaluation, and will be made publicly available together with the proposed model.

2 RELATED WORK

There have been multiple works that focus on manipulating highly deformable objects such as clothes. Most of these works concentrate on finding suitable grasping points either for towels [6, 14] or for more structured clothes like T-shirts, pants or sweaters [9, 15]. Typically, after capturing data with a depth sensor device, early methods concentrate on finding geometric cues [10, 12, 19] (i.e. cloth folds and wrinkles, cloth corners, etc.) or classify cloth deformation to indirectly infer the grasping points [8, 18]. However, these kinds of approaches are difficult to use for complex clothes, as the detected edges or other geometric cues lack semantic meaning, which needs to be compensated by using fiducial markers on the cloth [3] for a more reliable detection of grasping points.

Recent methods exploit the potential of neural networks. In order to train these networks, it is necessary to use a large amount of data, that can be achieved by means of generating synthetic datasets [5, 17]. Unfortunately, networks trained exclusively on synthetic data have problems generalizing when using real examples. For this reason, synthetic data is often mixed with real data which can be acquired by painting a white cloth with the desired annotation marks [6, 13]. This procedure is tedious and makes the data generation more complex as it involves robot manipulation to obtain images and pre-processing operations to extract the annotations. Therefore, other methods train the networks with synthetic data and later use a small number of real examples to fine-tune the grasping point detection network [11, 21]. In contrast, our work uses Domain Adaptation to narrow the gap between synthetic and real data. The main advantage of using DA is that it eliminates the need of collecting and annotating large real datasets, thus allowing the proposed network to be trained by supervision only

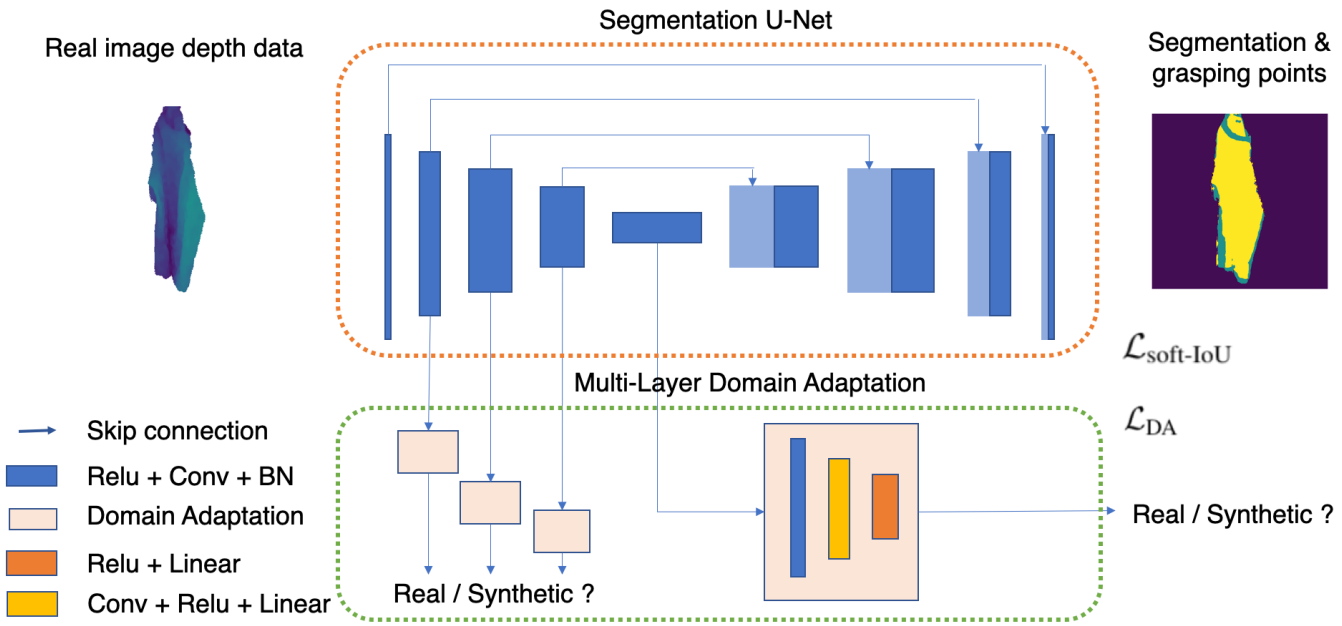

Figure 2: Our approach for fine-grained segmentation of cloth depth maps. It consists of two main modules, a U-Net (top) that segments the cloth parts and a multi-layered Domain Adaptation module (bottom) that helps to reduce the domain gap between real and synthetically generated depth maps. This strategy allows our network to generalize to real depth maps, despite being trained with synthetic supervision. The segmentation loss is only computed for the synthetic annotations, whereas unlabelled real data is leveraged in the multi-layer DA branch to reduce the gap between the real/synthetic features computed by the U-Net.

from synthetic examples, while achieving comparable results to the methods trained with real or a mixture of real and synthetic data. This characteristic makes our proposed method easily generalizable to various types of garments. Moreover, unlike the methods discussed above that focus on grasping point detection, our proposed approach is also able to detect semantically meaningful regions (e.g. neck, sleeve cuffs, hem) that can facilitate manipulating the cloth, especially in the case of occluded grasping points.

3 METHOD

In this section, we first formalize our problem, which is to segment depth maps of clothes into semantic regions that are tailored to perform manipulation and grasping tasks. We then describe the model as the training process we leverage to train only using synthetic labels.

3.1 Problem Formulation

Let $\mathbf{X}$ be a $H \times W$ depth map of a cloth hanging under gravity from a random point. Let us also define reference segmentation masks $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_C]$ for the C cloth labels, where $\mathbf{m}_i$ is a binary mask indicating the region in the depth map that belongs to the i-th label. As we shall see in the experimental section, we consider segments of very different sizes, from small regions defining grasping points, to elongated and large areas of semantic edges, as well as the whole body of the cloth. We will also show that treating the grasping points as regions, and detecting them using a segmentation approach, yields improved results compared to the methods that locate them using a network regressing directly their coordinates [21]. Furthermore, we define $\mathcal{S}^s = \{\mathbf{X}_i^s, \mathbf{M}_i^s\}$, $i = 1, \ldots, M_s$, a set of synthetically generated pairs of depth maps and ground truth masks, and $\mathcal{S}^r = \{\mathbf{X}_i^r, \mathbf{M}_i^r\}$, $i = 1, \ldots, M_r$, pairs of real depth maps and pseudo-ground truth masks, as described in Section 4.

Our goal is to estimate masks $\hat{\mathbf{m}}_i$ of relevant cloth parts from a given depth map $\mathbf{X}$, *i.e.* to learn the mapping $\mathcal{M} : \mathbf{X} \longrightarrow \hat{\mathbf{M}}$, where $\hat{\mathbf{M}} = [\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_C]$. In order to train $\mathcal{M}$, we explore a Domain Adaptation learning scheme that uses the synthetically generated ground truth data $\mathcal{S}^s$, as well as the real depth maps $\mathbf{X}_i^r$ without annotations. The real pseudo-ground truth masks $\mathbf{M}_i^r$ will not be used during the training of our main approach.

3.2 Model

The architecture used in our approach is illustrated in Fig. 2. It is composed of two main modules: a segmentation U-Net and multi-layer Domain Adaptation discriminators. Given an input depth map $\mathbf{X}$, the *segmentation U-Net* aims to classify every input pixel in $\mathbf{X}$ into one of the C cloth part categories. We implement this module following a standard convolutional U-Net architecture [16], with four encoder and four decoder blocks. As we show in the experimental section, training the U-Net network uniquely with synthetically generated data $\mathcal{S}^s$ does not generalize well to real depth maps. In order to narrow the domain gap between real and synthetic depth maps, we introduce a multi-layer adversarial-based Domain Adaptation strategy [4]. Following standard adversarial training [7], we use a two-step minimax optimization approach

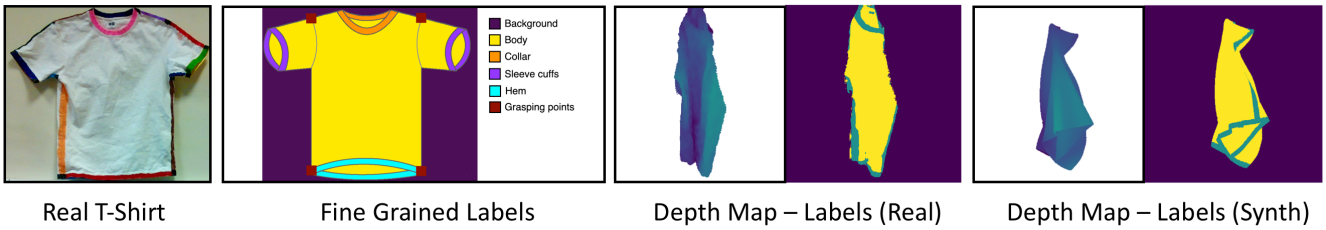

Figure 3: Left to right: T-shirt used for real domain data collection; Fine-grained labels present in our experiments; Depth map and segmentation labels for the real and synthetic domains (edges in *green*, main body in *yellow* and background in *dark blue*).

where we alternate the update of the U-Net and the discriminators. In the first step, we only update the U-Net to force the features extracted at each encoder block to confuse the multiple discriminators. Then, in the second step, we update the discriminator to distinguish between the feature maps from real ($\mathbf{X}_i^r$) or synthetic ($\mathbf{X}_i^s$) samples. This iterative process leads to the U-Net encoder layers to produce features that are indistinguishable by the discriminators, and thus leads to domain-independent features that bridge the gap between the synthetic and real domains. The domain discriminators we use are formed by convolutional layers followed by a fully connected layer that performs binary classification.

3.3 Loss Functions

Our loss function contains two terms, namely a semantic segmentation loss to assess the quality of the segmentation, and a domain-adaptation loss that forces the extracted features of the real and synthetic domains to be similar.

Segmentation loss. To optimize our segmentation model we maximize a weighted soft Intersection over Union (IoU) loss:

$$\mathcal{L}_{\text{soft-IoU}} = \frac{1}{|C|} \sum_c \frac{w_c \sum_x \tilde{\mathbf{M}}_c(x)\mathbf{m}_c(x)}{\sum_x \tilde{\mathbf{M}}_c(x) + \mathbf{m}_c(x) - \tilde{\mathbf{M}}_c(x)\mathbf{m}_c(x)}$$

where $\tilde{\mathbf{M}}_c(x)$ is the prediction score at the image location x for class c, and $\mathbf{m}_c(x)$ is the ground truth, which is a delta function in the correct class. Note that some classes occupy a large amount of area (*e.g.*, background), whereas some other classes (*e.g.*, edges) occupy a smaller area. To prevent the model from only focusing on the classes that occupy a larger extension, we added a class-specific weight factor w_c to reduce this label imbalance problem.

Domain Adaptation loss. Our DA loss $\mathcal{L}_{\text{DA}}$ is the standard iterative minimax loss using a binary cross-entropy loss with classes "real" or "synthetic" [7], where the U-Net aims to minimize this loss whereas the discriminators try to maximize it.

Total loss. We define the total loss as a linear combination of the two previous terms:

$$\mathcal{L} = -\mathcal{L}_{\text{soft-IoU}} + \alpha\mathcal{L}_{\text{DA}}\,, \tag{1}$$

where α is a hyper-parameter, and minimizing $-\mathcal{L}_{\text{soft-IoU}}$ is equivalent to maximizing $\mathcal{L}_{\text{soft-IoU}}$.

Our model is trained with both real and synthetic depth maps. However, the segmentation loss $\mathcal{L}_{\text{soft-IoU}}$ is only applied to the synthetic inputs $\mathbf{X}_i^s$, for which we have the ground truth labels $\mathbf{M}_i^s$. The real depth maps $\mathbf{X}_i^r$ are considered only in the DA loss $\mathcal{L}_{\text{DA}}$, as it does not require ground truth segmentation labels and allows us to leverage unannotated data.

4 DATASETS

We now describe the collection and annotation approach for our real and synthetic datasets.

4.1 Synthetic Domain

We show in Fig. 3 an example generated using the physics cloth engine from Blender [1]. The setup to generate the depth maps consists of a deformed T-shirt model surrounded by a rig of 36 cameras separated by steps of 10 degrees around the cloth. Specifically, the bounding box defined by the deformed mesh lies at the center of the circle, and we set the radius to 120 cm to ensure the whole T-shirt mesh is completely visible by all cameras. A 3D human body design suite [2] is used to obtain the T-shirt model. This model is defined by a quad mesh with 3500 vertices, which is the best topology for the cloth physics engine simulator. The cloth physics engine is based on a spring mass model, with several cloth fabric presets and several parameters that are tunable for adjusting the behaviour of the simulation. We use the *cotton* preset in the case of the T-shirt, and just modify the bending and stiffness parameters to achieve more realistic deformations. The T-shirt mesh is hung from a point and deformed by gravity. The deformation process is run for 250 steps on each physics simulation to ensure a rest position is achieved. Before running each simulation, the mesh is randomly rotated, and a vertex is also randomly chosen as a hanging point. We split the data into 5600/700/700 train/val/test samples. Note that the test samples come from different hanging configurations (hanging points) than the training samples. The images in consecutive frames are similar which would lead to a bias if random splits were used for training and testing. For the grasping point regression task, which will be described in the experimental section, we use 2737 training and 344 test samples as not all examples have visible grasping points.

We also carry out a normalization of the synthetic depth maps in the vertical and horizontal dimensions. For this purpose, the covariance matrices of the non-background pixels coordinates were averaged. The same matrix was obtained for the real dataset. The eigenvalues λ from the singular value decomposition were utilized to scale the synthetic images. Specifically, $\frac{\sqrt{\lambda^r}}{\sqrt{\lambda^s}}$ were used to scale the synthetic images along the vertical and horizontal axis to ensure a similar shape between real and synthetic depth maps.

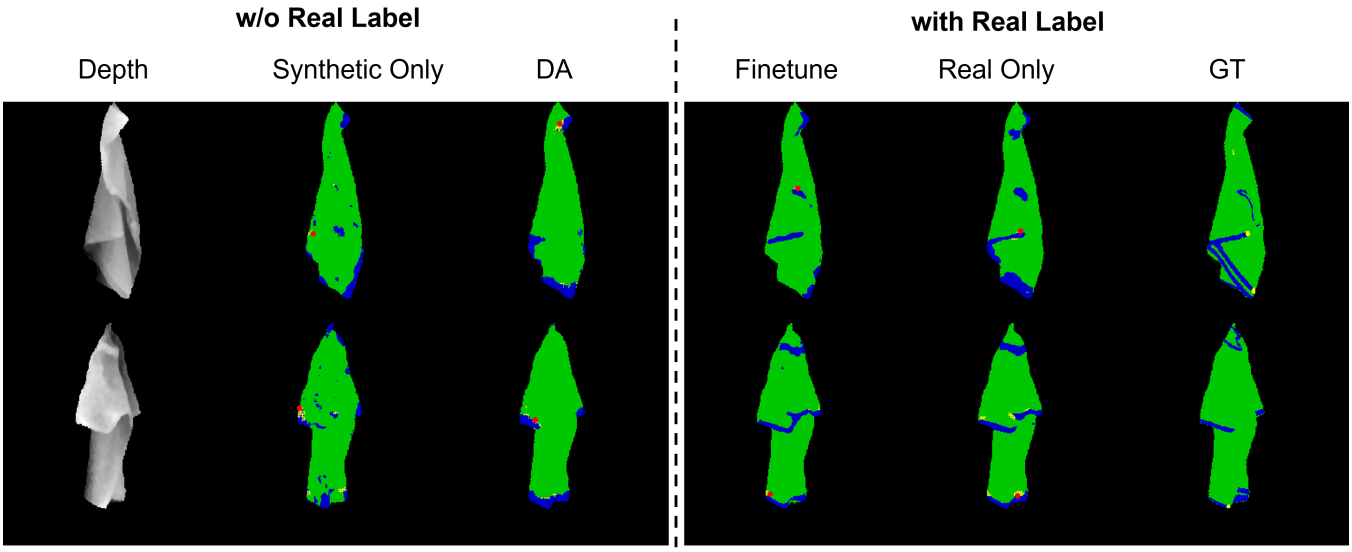

Figure 4: Qualitative results. Background, cloth body, edges are represented in *black, green, blue* respectively. Grasping points with highest confidence are highlighted in *red*.

4.2 Real Domain

To reduce the real data annotation burden, the T-shirt was first painted along the edges with different colors representing different edges and 4 grasping points as illustrated in Fig. 3-left. The T-shirt is grasped and hung by a Baxter robot. We manually adjusted the real-world setup to roughly match the appearance and dynamics of the simulation. Specifically, an Intel RealSense L515 camera is placed 120 cm away from the grasping point. The robot rotates the T-shirt every 10 degrees and depth images are captured. The pseudo labels annotation was done by training a U-Net segmentation model using RGB images and a small number of manually annotated examples with aggressive augmentation to adjust for the class imbalance. The generated pseudo-labels were afterwards verified by a human annotator.

Using this annotation process, we collect a real-domain dataset consisting of 504 samples with 6 fine-grained pseudo-labels, i.e., background, body, top, middle and bottom edges and grasping points. From these 504 examples, we use a 388/44/72 train/val/test split, where our 72 test samples come from two different hanging configurations, and only 48 samples include grasping point annotations. Since our DA approach does not need any real supervision, we further collect 1008 extra unannotated real depth samples which will be leveraged when training with DA. The annotations in the train set are only used in our finetuning experiments. All depth images were min-max normalized after removing the background with a threshold, and we use data augmentation techniques such as random horizontal flipping and cropping.

5 EXPERIMENTS

We now present the experimental results. First, we report the results for the segmentation of edges and grasping points, and then the results for fine-grained segmentation.

5.1 Main Results

The results for edge and grasping point segmentation are shown in Table 1 and Table 2. Note that except for *Synth/Synth*, the results reported refer to testing on real data for all other evaluated methods. The grasping point (GP) accuracy is measured as the median distance of the most confident prediction to the closest ground truth grasping point, which is reported both in pixels and cm. To match the GP pixel distance with the real-world scale, we measured the median length of the cloth length when it is hung on a hanger, which is 65 cm (230 pix). The performance for the background, body, and edges classes is measured with IoU.

For comparison, as baseline predictors of the grasping point we consider the centre of the cloth as well as a random grasping point. After the normalization of the synthetic data the median distance from the cloth centre to its closest grasping point is 11.7 cm (41 pix) for the synthetic dataset, and 12.7 cm (44.5 pix) for the real dataset. The median distance between randomly sampled points on the cloth to its closest grasping point is 10.4 cm (36.4 pix) for synthetic dataset, and 11 cm (39 pix) for real dataset. We evaluate only for one grasping point as in practice once the cloth is grasped by the predicted point, the model can be applied again to detect the next point.

Table 1 shows that generally the IoU for body segmentation is high, whilst the performance for the edges is significantly lower due to the complex folding of hanging cloth and overlapping of edges and body fabrics. As expected, due to the domain gap present, the performance when training on synthetic data (*Synth/Synth*) is higher than when training with synthetic data and testing with real data (*Synth/Real*), whilst finetuning with real data annotations increases the performance in the real data. On that note, our proposed DA method can bridge the domain gap without needing real data annotations, as shown in the increased performance of *DA* in Table 1 compared to the direct synthetic-to-real model (*Synth/Real*).

Table 1: Edge and grasping point results, where *GP dist* is the distance between the most confident predicted grasping point and the closest ground truth grasping point. We include the distance in pixels (image size is 256×256 pixels) and cm. Semantic segmentation performance is measured by Intersection over Union (IoU).

Train / Test	Background (IoU) ↑	Body (IoU) ↑	Edges (IoU) ↑	GP dist (pix) ↓	GP dist (cm) ↓
Synth / Synth	0.999	0.922	0.583	17.46	4.99
Synth / Real	0.998	0.757	0.097	38.49	11.00
Real / Real	0.998	0.919	0.289	45.55	13.01
Finetune	0.997	0.894	0.328	32.38	9.25
DA + Finetune	0.999	0.891	0.276	22.62	6.46
DA	0.998	0.852	0.209	25.29	7.23

Table 2: Fine-grained results for 6 class labels. The performance for the edges drops as the model struggles to discriminate between top, bottom, and side edges. The grasping point error (*GP dist*) of *DA* is nearly half of that when training with synthetic or real data only.

Train / Test	Backgr. (IoU) ↑	Body (IoU) ↑	Top (IoU) ↑	Middle (IoU) ↑	Bottom (IoU) ↑	GP dist (cm) ↓
Synth / Synth	0.999	0.929	0.278	0.560	0.567	3.89
Synth / Real	0.997	0.861	0.015	0.129	0.040	13.36
Real / Real	0.997	0.916	0.025	0.130	0.108	15.78
Finetune	0.998	0.897	0.041	0.200	0.174	8.56
DA + Finetune	0.998	0.917	0.019	0.313	0.294	6.67
DA	0.997	0.846	0.050	0.307	0.138	7.05

Specifically, we note that our DA approach reduces the grasping point prediction error by 34.3% compared to *Synth/Real* without using any real data annotation. Additionally, we can couple our *DA* method with the available real annotations to further improve the performance, as shown in *DA + Finetune*.

In Table 2 we report our fine-grained segmentation results, including additional top, middle, bottom labels for edges as well as grasping points. The performance for the edge-related classes drops compared to Table 1 as the model struggles to discriminate between bottom, middle and top edges. Regarding grasping points, similarly to what is observed in Table 1, the results for *Real/Real* are lower than *Synth/Real* due to the limited size of the real dataset, whereas finetuning with real data greatly improves the quantitative performance. More importantly, the grasping points obtained by our proposed DA are almost twice as accurate as those obtained with *Real/Real* or *Synth/Real*.

5.2 Additional Results

Our proposed segmentation approach also achieves good performance compared to the state-of-the-art grasping point regression methods. Due to inaccessibility to other datasets and implementations, we implemented a regression network based on the architecture proposed in [21], which was trained with the same data used when training our method.

Table 3: Comparison of different methods for grasping point prediction task. Results reported are distance in cm. In *2 class segmentation*, each pixel is predicted to be GP / not GP. *4 class segmentation* and *6 class segmentation* are the results from Table 1 and Table 2.

Method	Synth/Synth	Synth/Real	DA
Regression method [21]	10.25	13.90	-
2 class segmentation	8.57	9.71	8.32
4 class segmentation	4.99	11.00	7.23
6 class segmentation	3.89	13.36	7.05

Table 3 shows that approaches based on segmenting the depth maps reach higher performance in the task of grasping point prediction than the regression approach proposed in past works [21]. The performance when training and testing in synthetic data (*Synth/Synth*) shows the increased achievable performance with our segmentation-based approach compared to a regression-based method. Furthermore, when fully-training with synthetic data and testing on real data (*Synth/Real*) adding more semantic classes reduces the performance, whereas that trend is reversed when using DA approach. These results highlight the benefits of our multi-layer DA method as it allows us to leverage the synthetic fine-grained annotations to improve the grasping point prediction performance.

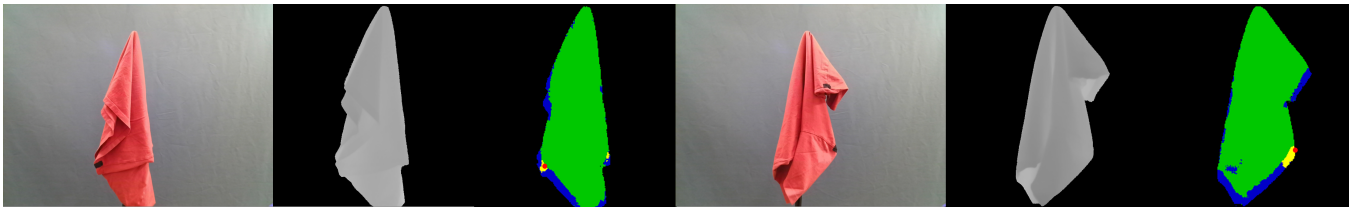

Figure 5: Visualization of the results with a different depth camera. Background, cloth body, edges are denoted in *black*, *green*, *blue* respectively. Grasping points with highest confidence are highlighted by dots in *red*.

Finally, we show qualitative results in Fig. 4 for our model trained without (left) and with (right) the real-domain labels. Finetuning with real labels visibly improves the accuracy of the segmentation, which is especially notable in the increased ability of detecting edges. In general, leveraging real data allows the model to discriminate finer regions compared to the model trained only with synthetic data. Our DA-based method also predicts the edges reliably, as DA aids the model to match the distribution of the features in both domains and thus improves the ability of the model to generalize.

To test the generalization ability of our proposed DA approach, we collected an additional real dataset with 25 samples using a different T-shirt and depth sensor (Azure Kinect). We apply our 4-class method to that dataset without any further training, obtaining a median grasping distance of 13.44 cm. The obtained performance is better than the median distance from the cloth centre to its closest grasping point (15.70 cm), which indicates a good generalization ability. Qualitative results are given in Fig. 5.

6 CONCLUSIONS

In this paper, we have investigated the segmentation of depth maps from highly deformed clothes into semantic regions that are useful for subsequent downstream tasks such as cloth grasping manipulation for folding or assisted dressing. We made two main contributions. First, the proposed architecture allows predicting regions of different types and extents, from local grasping points to larger semantic edges, without the need for retraining. Second, we devise a learning methodology that makes it possible to train the network using only semantic annotations on synthetic data, and addresses the domain gap between real and synthetic depth maps via a multi-layered Domain Adaptation strategy. The experiments show promising results, where coupling our Domain Adaptation approach with only synthetic ground truth annotations allows us to obtain results on par with a network trained with real data without needing to annotate the real images.

REFERENCES

[1] [n. d.]. Blender. https://www.blender.org. Accessed: 2021-09-14.

[2] [n. d.]. Makehuman. http://www.makehumancommunity.org. Accessed: 2021-09-14.

[3] Christian Bersch, B. Pitzer, and S. Kammel. 2011. Bimanual robotic cloth manipulation for laundry folding. *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2011), 1413–1419.

[4] Ozan Ciga, Jianan Chen, and Anne Martel. 2019. Multi-layer domain adaptation for deep convolutional networks. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 20–27.

[5] Enric Corona, Guillem Alenyà, Antonio Gabas, and Carme Torras. 2018. Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition* 74 (2018), 629–641. https://doi.org/10.1016/j.patcog.2017.09.042

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

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[8] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. 2009. Clothes state recognition using 3D observed data. In *2009 IEEE International Conference on Robotics and Automation*.

[9] Yinxiao Li, Danfei Xu, Yonghao Yue, Yan Wang, Shih-Fu Chang, Eitan Grinspun, and Peter K. Allen. 2015. Regrasping and Unfolding of Garments Using Predictive Thin Shell Modeling. In *2015 IEEE International Conference on Robotics and Automation*.

[10] Jeremy Maitin-shepard, Marco Cusumano-towner, Jinna Lei, and Pieter Abbeel. 2010. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *In International Conference on Robotics and Automation (ICRA*.

[11] Ioannis Mariolis, Georgia Peleka, Andreas Kargakos, and Sotiris Malassiotis. 2015. Pose and category recognition of highly deformable objects using deep learning. In *2015 International Conference on Advanced Robotics (ICAR)*. 655–662. https://doi.org/10.1109/ICAR.2015.7251526

[12] Luz María Martínez and Javier" Ruiz-del Solar. 2018. Recognition of Grasp Points for Clothes Manipulation Under Unconstrained Conditions. In *RoboCup 2017: Robot World Cup XXI*. 350–362.

[13] Jianing Qian, Thomas Weng, Brian Okorn, and L. Zhang. 2020. Cloth Region Segmentation for Robust Grasp Selection. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), 9553–9560.

[14] Arnau Ramisa, G. Alenyà, F. Moreno-Noguer, and C. Torras. 2011. Determining Where to Grasp Cloth Using Depth Information. In *CCIA*.

[15] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. 2014. Learning RGB-D descriptors of garment parts for informed robot grasping. *Eng. Appl. Artif. Intell.* 35 (2014), 246–258.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[17] Krati Saxena and Tomohiro Shibata. 2019. Garment Recognition and Grasping Point Detection for Clothing Assistance Task using Deep Learning. In *2019 IEEE/SICE International Symposium on System Integration (SII)*. 632–637.

[18] Jan Stria and V. Hlavác. 2018. Classification of Hanging Garments Using Learned Features Extracted from 3D Point Clouds. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), 5307–5312.

[19] Li Sun, Gerardo Aragon-Camarasa, Simon Rogers, and J. Paul Siebert. 2015. 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)*. 185–192.

[20] Ulrich Viereck, Andreas Pas, Kate Saenko, and Robert Platt. 2017. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In *Conference on robot learning*. PMLR, 291–300.

[21] Fan Zhang and Yiannis Demiris. 2020. Learning Grasping Points for Garment Manipulation in Robot-Assisted Dressing. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*.

[22] Xiaoshuai Zhang, Rui Chen, Fanbo Xiang, Yuzhe Qin, Jiayuan Gu, Zhan Ling, Minghua Liu, Peiyu Zeng, Songfang Han, Zhiao Huang, et al. 2022. Close the Visual Domain Gap by Physics-Grounded Active Stereovision Depth Sensor Simulation. *arXiv preprint arXiv:2201.11924* (2022).