APPENDIX A
METHOD OVERVIEW

Fig. A is a pipeline diagram for our method, which comprises of cloth region segmentation, grasp selection, and grasp execution.

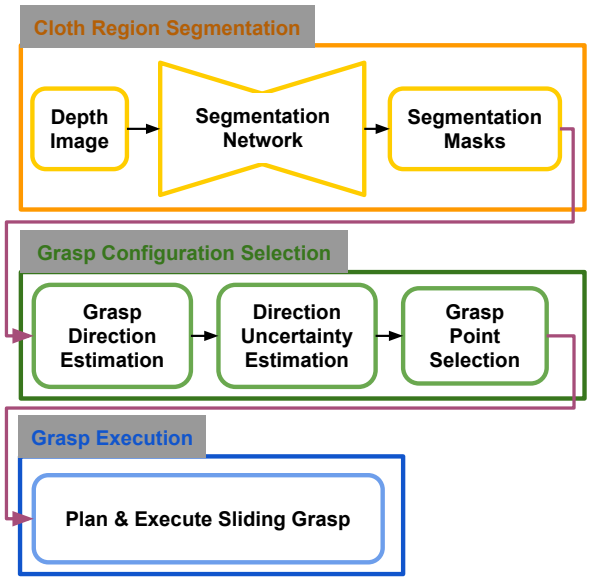


Fig. A: Pipeline for our method. Cloth region segmentation takes a depth image and outputs segmentation masks for cloth edges and corners. Grasp selection uses the masks to compute a grasp point and direction in the camera frame. Grasp execution transforms the grasp configuration into the robot frame and executes the grasp.

APPENDIX B
TRAINING DETAILS

The network was implemented in PyTorch [13] and trained using the Adam optimizer [9]. We augmented the data during training with random image flips and rotations to improve robustness. All training was performed on an Ubuntu 16.04 machine with an NVIDIA GTX 1080 Ti GPU, a 2.1 GHz Intel Xeon CPU, and 32 GB RAM.

Hyperparameters used for training the cloth segmentation network

- Learning rate: 1e-05
- Batch size: 64
- Number of rotation augmentations: 32
- Loss: Binary cross-entropy

APPENDIX C
GRASP EXECUTION

This section contains details about calculating grasp execution given the grasp configuration output from the network. The configuration $(\mathbf{p}, \alpha)$ specifies the grasp point on the cloth and the direction for the sliding grasp. This configuration is specified in image coordinates; to transform it into the world frame, we perform a 2D-to-3D projection using known camera intrinsics and extrinsics. This provides an intermediate 6D pre-grasp pose $\tilde{\mathbf{g}}$ consisting of the 3D position of the target cloth point (corresponding to $\mathbf{p}$ in 2D), and the 3D orientation of the end-effector (corresponding to α in 2D). The intermediate pre-grasp pose $\tilde{\mathbf{g}}$ is oriented top-down and rotated about the z-axis in the world frame. We apply a final transformation that tilts the grasp pose about the horizontal x-axis by 45-degrees to obtain a new pre-grasp pose $\mathbf{g}$. This pose allows one of the fingertips to get under the cloth during the slide action. This transformation also includes a z-offset to account for the z-height of the gripper tip lowering due to the rotation. Finally, we compute offsets to $\mathbf{g}$ in the xy plane parallel to the workspace to get pre-slide and post-slide poses. As shown in Fig. 4, the sliding grasp policy moves to the pre-slide pose, translates to the post-slide pose, then pinches to grasp the cloth.

The sliding grasp policy was implemented for the robot using off-the-shelf MoveIt! software [3]. The default fingertips of the Weiss gripper were too thick to get under the cloth during the sliding maneuver, so we 3D-printed and attached thinner fingertips.

APPENDIX D
EXPERIMENTAL RESULTS

In this section we provide additional explanation for why the baselines performed poorly, along with an ablation experiment for our method.

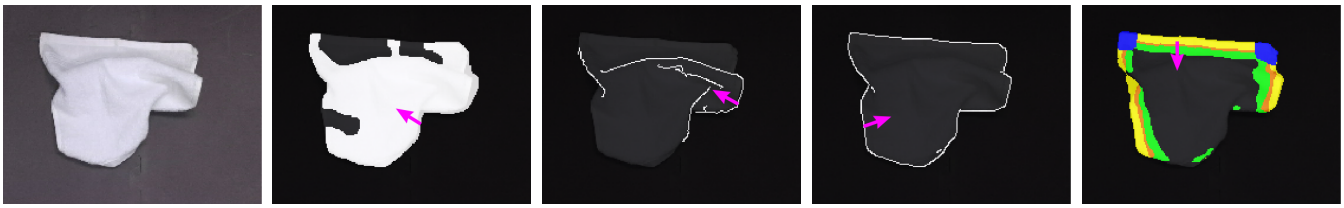
A. *Edge Baselines*

Fig. B visualizes the different edge grasping methods.

Canny-Depth relies on the intensity of depth gradients to find cloth edges, but depth gradients occur for both cloth edges and large folds. Segment-Edge fails due to noisy segmentation; because the cloth is thin, parts of the cloth can fall within the inlier threshold of the RANSAC table segmentation, despite careful parameter tuning. Still, even with a clean segmentation, grasping at an edge point on the segmentation mask often results in grasping a cloth fold for our highly crumpled cloth configurations. Canny-Color uses color gradients to find edges. It is less affected by noise compared to the depth-based baselines, as the white cloth stands out from the darker background of the table, resulting in strong edges. However, this method is still unable to discriminate between real cloth edges from folds, resulting in failure in a majority of grasp attempts.

B. *Corner Baselines*

The Harris-Depth baseline performs poorly because it looks for large changes in the gradient in all directions, which could result in false positives instead of real corners. Most of the grasp point selections from this baseline were on wrinkles and folds than on the cloth. The Harris-Color baseline performs better than depth, possibly because there are fewer false positives given the white on black input images. White cloth corners against the darker workspace surface can be easily detected; however, corners lying on top of the cloth are less



(a) Cloth Pose (for reference). (b) Segment-Edge. (c) Canny-Depth [2]. (d) Canny-Color [2]. (e) Our Method.

Fig. B: Our method correctly identifies most of the apparent edges of the cloth as folds, whereas the other methods fail to make this distinction. (b)-(e) visualize the output of each method on top of the reference image (a). Note that the color image is only provided as input to Canny-Color (d); all other methods take the corresponding depth image as input.

TABLE II: Ablations on Grasping Cloth Edges

Method	Grasp Success
No-Direction-Prediction	0.2
No-Directional-Uncertainty	0.4
Our Method	$\mathbf{0.7 \pm 0.20}$

1 trial per ablation, 10 grasp attempts in trial

likely to be detected. For our difficult randomly crumpled cloth configurations, the corners are not always cleanly visible against the surface, and often lie in configurations that are difficult to discriminate in 2D.

C. Ablations

We perform ablations on our method to determine the relative contribution of the different components of our method to grasp success. Our full method consists of segmenting cloth regions using a neural network, determining the grasp direction for all segmented edge/corner pixels using their nearest segmented inner edge pixels, and selecting a grasp point with the lowest grasp directional uncertainty.

We perform the following ablations of our method:

- "No-Direction-Prediction" still uses the cloth segmentation network. However, rather than determining the grasp direction using the our method, this ablation determines the grasp direction by fitting a bounding box around the segmented outer edge pixels and setting the direction to be the vector pointing to the center of the box. Instead of using the point with minimum directional uncertainty, it randomly selects the grasp point from the set of outer edge pixels.
- "No-Directional-Uncertainty" still uses the cloth segmentation network of as well as our method for determining the grasp direction. However, rather than computing the grasp directional uncertainty to choose a grasp point, this ablation chooses a grasp point randomly.

The results are shown in Table II. The ablations underperform the full method, demonstrating that our method for estimating the grasp direction as well as our method for estimating directional uncertainty help to choose more robust grasps. We observe No-Direction-Prediction selecting grasp directions near-parallel to real edges instead of orthogonally, because it always chooses directions toward the center of the segmentation bounding box. The performance of No-Directional-Uncertainty vs. No-Direction-Prediction provides evidence that using the inner edge segmentation to determine the grasp direction improves grasp success. Comparing our full method with No-Directional-Uncertainty shows that selecting the grasp point with minimal directional uncertainty outperforms random grasp point selection.